Kernel-Based Classification of Protein Structure and Function from Amino Acid Sequences

A thesis submitted for the degree of Doctor of Philosophy

by

Jonathan James Ward

University College London

April 29, 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

To Kristina

Abstract

The thesis describes the application of kernel methods and, in particular, the support vector machine (SVM) to the classification of protein structure and function. The thesis is divided into two related halves with chapters 2 and 3 containing descriptions of methods for predicting different aspects of protein structure. Chapter 4 investigates the functions of disorder in the proteome of a model eukaryote and Chapter 5 describes algorithms and data sources for inferring protein function. The data sources include structure predictions and other properties that can be derived directly from amino acid sequences.

Chapter 2 describes a new method for the prediction of secondary structure using an SVM learning algorithm. This is presented as a guide to the application of SVMs to problems in bioinformatics, and includes a discussion of the positive and negative aspects of the technique. The final prediction method is shown to have comparable performance to several of the most accurate modern methods.

The third chapter discusses the development of a method to recognize native disorder from amino acid sequences. This predictor (DISOPRED2) is shown to be the most accurate contemporary method on targets from the fifth CASP experiment. The false positive rate of DISOPRED2 is determined using ordered structures from the Protein Data Bank, and the classifier is then used to estimate the frequency of disorder in complete genomes. The final part of this chapter presents the design and implementation of a publicly-available web service for disorder prediction.

The fourth chapter describes the use of DISOPRED2 to investigate the functional annotations that are associated with long predictions of disorder in the proteome

of the model organism *Saccharomyces cerevisiae*. This chapter also provides several biochemical and evolutionary explanations for the disparity in the predicted frequencies of disorder between eukaryote and prokaryote proteomes. The chapter also demonstrates that the boundaries between structural domain have a propensity toward being predicted as disordered by DISOPRED2.

The final research chapter discusses the development of machine learning methods for determining the function of unannotated proteins. Individual classifiers are trained using phylogenetic profiles, structure predictions and simple features derived from the amino acid sequence to predict the function of yeast proteins in the absence of significant sequence similarity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The genome sequencing projects have led to a massive proliferation of data that represent most of the information required to construct and maintain living cells in an increasing number of organisms, which range in complexity from viruses and the simplest bacteria to the higher eukaryotes (Venter et al., 2001; Lander et al., 2001; Venter et al., 2004; Kyrpides, 1999). The genome of a particular organism contains only the DNA sequence, which acts as a passive information source encoding proteins and RNA. The challenge ahead is to understand how this code, and the functional gene products that arise from it, co-ordinate the development and maintenance of living cells. This thesis describes the development of several techniques that provide limited solutions to two of the most important outstanding problems of the early part of the post-genomic age; inferring the structure and function of proteins from their amino acid sequences.

The discipline of bioinformatics has developed over the past few decades to organize and gain greater understanding of the vast quantity of data arising directly and indirectly from the genome sequencing projects. Most research in bioinformatics is directed at developing tools that can generate biological information more quickly or efficiently than is possible by experiment. The other objective of research in bioinformatics is to organize and use (typically large) sources of data to provide a more general perspective on biological systems. The majority of this thesis is concerned with the first of these objectives, and specifically to inferring the mapping between the properties of an amino acid sequence and its structure or function. This is found by *learning* an approximation to the mapping between sequence and structure or function from the large number of amino acid sequences with solved structures or known function. This mapping can then be applied to unannotated amino acid sequences to *predict* the likely structure or function. In general, it is expected that these predictions will guide future experiments, although very high fidelity predictions may eventually obviate the need for further experimentation.

The field of *supervised learning* is concerned with the development of algorithms that generate a mapping from an input space (in this case residues or entire amino acid sequences) to the output domain (local structure or protein function class) from a finite set of training examples. The key objective of most supervised learning algorithms is to achieve the lowest possible error on unseen examples, a property also known as *generalization.* Chapters 2, 3 and 5 provide comparisons of several supervised learning algorithms. These include the feed-forward neural network, trained using variants of the back-propagation algorithm, and the support vector machine (Cristianini and Shawe-Taylor, 2000).

One of the potential advantages of the support vector machine (SVM) is that it is designed to optimize the accuracy on unseen examples using a criterion suggested by statistical learning theory (Vapnik, 1998), whereas the neural networks are trained using an *ad hoc* approach to improving generalization. The SVM is also one implementation of a particular class of learning algorithms that can be trained using a matrix of inner products between a set of Euclidean feature vectors. These *linear* learning machines can be generalised to non-linear problems by using a *kernel* function, which defines a mapping of the example vectors to an implicit (usually high-dimensional) feature space.

The other general objective of bioinformatics or computational biology is to provide insight into biological systems that cannot be obtained by experiment alone. Most research in experimental biochemistry, molecular genetics and cell biology is necessarily reductionist, and involves focusing on a very specific system. Computational techniques such as databases, machine learning, distributed computing and numerical modelling are necessary to integrate these disparate, local information sources into a global view of biological systems. The large-scale experimental technologies that have accompanied the genome projects, such as microarrays (Shipp et al., 2002; Brazma and Vilo, 2000), structural genomics (Burley, 2000) and 2-

hybrid screens (Uetz et al., 2000) have also necessitated the development of automatic methods for processing the resultant data (Alter et al., 2000; Bray et al., 2004; Jansen et al., 2003).

As of December 2004, the complete genome sequences of a total of 235 organisms were available, in addition to the 945 complete viral genomes (Kyrpides, 1999; Kersey et al., 2003). This figure is set to continue its dramatic dramatic increase as a result of the 1232 ongoing initiatives to sequence 538 prokaryote and 463 eukaryote genomes. Indeed, the sequencing of bacterial genomes is now considered routine, to the extent that it is now possible to begin mapping the genetic diversity of entire bacterial ecosystems, such as the marine environment of the Sargasso Sea (Venter et al., 2004). The rapid accumulation of sequence data and the slow and labourious nature of inferring protein structure and function experimentally would appear to suggest that it would be many decades before a complete picture of biological systems would begin to emerge (see Figure 1.1).

However, Nature generally evolves new functions by adapting a pre-existing sequence rather than *de novo* generation of a completely new sequence. This has the consequence that a high degree of similarity between two sequences implies that the two proteins adopt similar structures and that they have related functions. Therefore, although the total of around 10^8 sequenced proteins will continue to grow, there may be as few as one thousand naturally-occurring globular folds (Wolf et al., 2000). The structure and function of a novel protein can therefore be inferred with high fidelity, if it has high sequence similarity with another protein that has been annotated experimentally. Sequence searching techniques are therefore the standard tools for annotating protein sequences (Krogh et al., 1994; Altschul et al., 1997).

Nature's parsimonious adaptation of pre-existing structures can also be used to provide clues to the likely structure and function of proteins that do not share similar

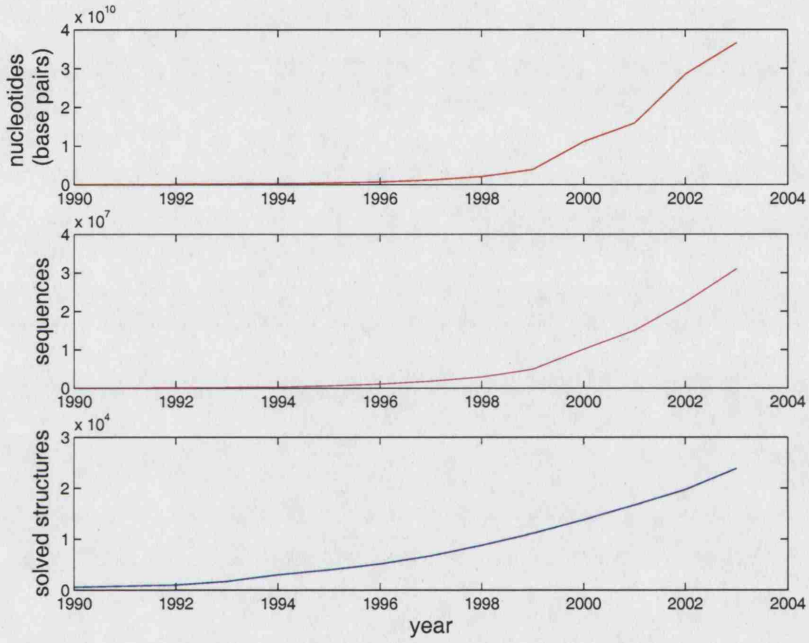

Figure 1.1: Growth in the number of nucleotides and gene products stored in the Genbank database. The growth in the number of structures is taken from the Protein Data Bank (PDB), which is an international repository of experimentally-determined protein structures (Berman et al., 2000). The number of solved structures in the PDB is currently growing by approximately 4000 structures per year. Genbank's gene product database is increasing in size at rate of around 8.7×10^6 sequences per year (Benson et al., 2004).

sequences with proteins in the existing knowledge base. For example, certain short sequence patterns, known as *motifs*, are used in numerous sequences for performing a specific biochemical activity such as modifying enzyme activity by glycosylation or metal binding. Sequence searching can also be used to determine the *conservation* of residues within the target sequence when compared with other members of the family. This information is used by the most successful methods for predicting protein structure (Jones, 1999; Ward et al., 2004b). Sequence search techniques can also be used to detect other evolutionary patterns that are indicative of protein function such as gene duplications, fusions and orthology (Zhang et al., 2003; Marcotte et al., 1999a; Pellegrini et al., 1999).

The remainder of this introduction describes the techniques that are combined in subsequent chapters to predict the structure and function of amino acid sequences. This begins with a description of algorithms for performing sequence searches with particular emphasis on the PSI-BLAST method (Altschul et al., 1997; Durbin et al., 1998), as this program is used throughout the thesis. The second half describes various types of neural network and a short derivation of the SVM and its theoretical justification. The thesis is structured so that reviews of the relevant literature are presented in the introduction to each research chapter. This allows the biological implications of the research and its contribution to bioinformatics to be assessed in the context of wider research in the field. The overall discussion in Chapter 6 includes a brief summary of the major findings of the thesis, and also discusses the research question that is common to all chapters of this thesis with the exception of Chapter 4; under what conditions support vector machines are likely to outperform other supervised learning algorithms.

1.1 Techniques for Comparing Sequences

The techniques from bioinformatics that have had greatest impact on the wider scientific community are algorithms for the comparison of amino acid and nucleic acid sequences. The culmination of these efforts are the BLAST (Basic Local Alignment Search Tool) and PSI-BLAST (Position-Specific Iterated BLAST) algorithms, which are amongst the most widely used in science (Altschul et al., 1997, 1990). The basic premise behind sequence comparison is that the greater the similarity between protein or DNA sequences, the more recent the divergence from a common ancestor and the more structural and functional characteristics will be shared by the two sequences.

The early techniques for sequence comparison had the aim of constructing optimal alignments between DNA or protein sequences according to a matrix of scores for matching letters in the respective alphabets. The alignment of sequences is often a preliminary step in the construction of phylogenetic trees and other evolutionary studies of proteins (Graur and Li, 2000; Durbin et al., 1998). Multiple sequence alignments also indicate which residues are conserved in closely-related sequences, and this information can be used to infer structural (Rost, 1996) and functional properties of the protein (Todd et al., 2001).

The two most important developments that preceded BLAST were the Needleman and Wunsch (1970) dynamic programming algorithm for obtaining an optimal global alignment of two sequences and the Smith and Waterman (1981) modification to the score matrix for obtaining the optimal local alignment. This was followed by the Karlin and Altschul (1990) method for calculating rigourous estimates for the statistical significance of ungapped local alignment scores. A major limitation of the Smith and Waterman (1981) local alignment algorithm is that the time and space complexity are both $O(mn)$, where m and n are the lengths of the two sequences in

the comparison. This problem is particularly acute when the search is carried out between a single sequence $O(m) = 10^2$ and a large sequence database $O(n) = 10^{10}$. This computational complexity was overcome by developing heuristics for detecting sequences within the database that are likely to have a high-scoring alignment with the query sequence.

The first heuristic search methods FASTA and FASTP identified 'words'[1] in the query sequence that were identical to other words in the sequence database (Pearson and Lipman, 1988; Pearson, 1990). The rationale for using word matches is that a high-scoring local alignment between the query and a sequence in the database is likely to contain at least one pair of matching words. A high density of word matches along the diagonal of the alignment matrix indicates a high scoring local alignment and triggers extension of the local alignment.

The original version of BLAST used a slightly improved search heuristic but its major advantage over FASTA was the use of the Karlin and Altschul (1990) estimates of the statistical significance of local alignments. The BLAST search heuristic identifies words[2] in the query sequence with a match score above a particular threshold. The sequence database is then scanned for the high-scoring word hits, with detection of a hit triggering an alignment to be extended in both directions (Altschul et al., 1990). If the alignment score is above a threshold, the Karlin and Altschul (1990) statistics are used to calculate the probability of recovering a better alignment between a query sequence of length m and a sequence database of length n under the null model[3].

The PSI-BLAST algorithm improved the efficiency over BLAST by only extending local alignments between two high-scoring word hits that are within close

[1] k-mers of a particular length.

[2] The default for searching amino acid sequences is $k = 3$ in PSI-BLAST version 2.

[3] Zeroth order Markov model; amino acids occur independently.

proximity of each other. The detection of remote homologues was also increased by allowing gaps in the local alignment for the gapped-BLAST method, and particularly by the construction of profiles for subsequent PSI-BLAST searches. The profile or position-specific scoring matrix (PSSM) is constructed by performing a multiple alignment of the sequences recovered from the initial BLAST search. The frequencies of the amino acids at each position in the multiple alignment are then used to weight the original scoring matrix to account for the residues that are present in the family of sequences recovered from the initial search. This PSSM is then used to detect more remote homologues in the subsequent BLAST search, with the process repeated until convergence[4] or the user-specified number of search rounds has been completed. PSI-BLAST's sensitivity, computational efficiency, solid statistical foundation and easy usage have made it a standard tool in many branches of biochemistry and genetics.

PSI-BLAST is also used throughout this thesis with the PSSMs used successfully for the prediction of protein structure (Ward et al. (2003, 2004b), Chapters 2 and 3). The annotation of protein function shown in the final chapter also makes use of PSI-BLAST for estimating the accuracy of function assignment using homology and for generating phylogenetic profiles for *ab initio* prediction of protein function. PSI-BLAST is also used in the detection of gene fusion events, which are a specific type of second-order interaction in the graph of homology relationships between genomes.

Although this thesis is also concerned with investigating the evolutionary and physical properties of proteins, a second theme is the comparison of support vector machines with other classification algorithms. The SVM is a relatively new development, and has inspired much recent work in machine learning. SVMs have some known advantages over the feed-forward neural network, which currently prevails

[4]No hits are recovered with similarity scores above the threshold for inclusion in subsequent PSSMs.

as the most widely used classification algorithm in bioinformatics. The comparison of the SVM with other supervised learning algorithms is carried out to establish the conditions for the SVM producing greater accuracy than neural networks and other techniques. The following introduction to support vector machines and the discussion in Chapter 5 also describe the advantages of using kernels for encoding non-Euclidean sources of data such as graphs, trees and sequences.

1.2 Support Vector Machines

The central goal of *supervised learning* is to construct a mapping or hypothesis that classifies unknown examples with the least error, a property known as *generalization.* The support vector machine, developed by Vapnik and co-workers (Vapnik, 1998), is designed to optimise generalization and has become established as a standard technique in numerous pattern recognition applications, which include the detection of translation-initiation sites in DNA sequences (Zien et al., 2000) and the classification of microarray expression profiles according to gene functional classes (Brown et al., 2000) and tissue types (Furey et al., 2000).

Machine learning is often applied to problems where it is difficult to construct a physical model of the processes generating the data but where a reasonable number of examples are available. A typical example is the recognition of isolated handwritten digits. The basic measurements associated with a particular example, such as pixel values, are often referred to as *attributes* with higher-level characteristics such as character height and width described as *features* (Bishop, 1995). This distinction does not exist for pattern recognition problems where the properties or measurements that best represent each example are not obvious, as is usually the case in bioinformatics. The two terms will therefore be used interchangeably in the rest of this thesis. The selection of features or attributes for a particular classification task

often impacts greatly on classifier accuracy, as has been shown previously in the field of secondary structure prediction (Rost and Sander, 1993; Jones, 1999) and again in Chapter 3 for the prediction of flexible protein structures.

The solution of a supervised learning problem is chosen from a set of candidate functions which map from input space to output domain. These are normally restricted to a set of functions within the hypothesis space. The learning algorithm uses a criterion called a *learning bias* for selecting a suitable model of the training data. For example, an appropriate bias for approximating any arbitrary function would be to select the hypothesis that minimizes the mean squared error between the model outputs and the target values. Support vector machines incorporate a learning bias, founded in frequentist learning theory, that is designed to optimise the generalization of the resulting classifier. It is believed that the learning bias is one of the reasons why SVM classifiers have prediction accuracies that often match or exceed those of other techniques. SVMs also avoid some of the disadvantages of other learning algorithms, including the curse of dimensionality and overfitting, local minima and the setting of many tunable parameters (Cristianini and Shawe-Taylor, 2000).

1.2.1 Introduction to Probably Approximately Correct (PAC) Learning

Theoretical machine learning is concerned mainly with constructing classifiers that perform well on unseen examples. The greatest theoretical obstacle to achieving generalization is overfitting, which also appears in the guises of the bias-variance trade-off and capacity control (Bishop, 1995; Vapnik, 1998). A model chosen from too rich a set of hypotheses (overcapacity, high variance) can lead to fitting of the noise on the training examples whilst a model chosen from a set of hypotheses that is

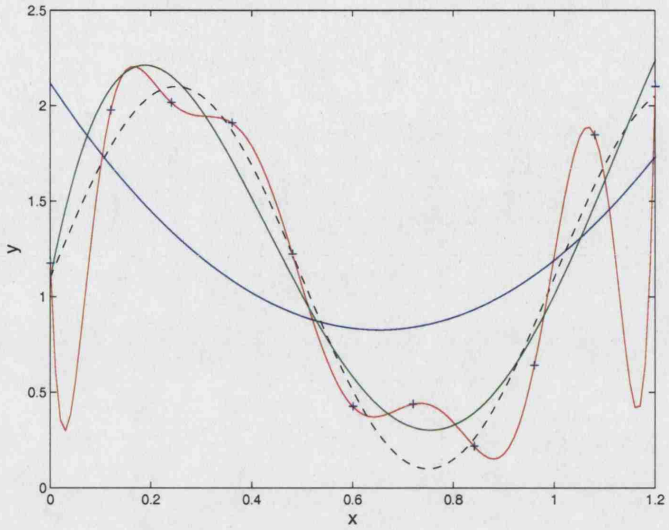

Figure 1.2: An illustration of over- and under-capacity. Polynomial fits of the function $y = \sin(2\pi x) + 1.1$ (shown by dotted black curve). The data points (crosses) have been generated from the function with the addition of Gaussian noise term with a mean of zero and variance of 0.15. The blue, green and red curves show second, fourth and tenth order polynomial fits that optimise the sum-of-squares error. The blue curve has insufficient free parameters to model the data (undercapacity) and the red curve fits the points with zero error but is a very poor approximation to the underlying function (overcapacity).

too restrictive (undercapacity, high bias) may not be able to represent the properties of the inputs that distinguish between classes (see Figure 1.2). This problem has been addressed by several approaches including information theory and Bayesian statistics (Mackay, 2003).

The motivation for support vector machines is derived from classical or frequentist statistics and the probably approximate correct (PAC) learning methodology (Vapnik, 1998). The key assumption is that the data used in training and testing are generated independently and identically from some unknown but fixed distribution. The measure of error for binary classification is the probability that a randomly

generated example $\mathbf{x}$, which is usually a vector in Euclidean space $\mathbf{x} = (x_1, \ldots, x_n)$ with a label $y \in \{1, -1\}$ is misclassified

$$R = \int \frac{1}{2}|y - f(\mathbf{x})|p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y \tag{1.1}$$

where $f(\mathbf{x}) \in \{1, -1\}$ is the classifier output and $p(\mathbf{x}, y)$ is the probability distribution on $X \times \{-1, 1\}$ where X is the set of all possible input vectors.

This expected error or risk measures the generalization of the learning machine. However, this definition is rarely of practical use since estimating the density, $p(\mathbf{x}, y)$, is usually more difficult than simply discriminating between the classes (Vapnik, 1998). The aim of PAC analysis is to construct upper bounds on the expected error in terms of quantities that describe the classifier and the learning problem. These bounds, which hold with a certain probability, can then be used to select the hypothesis that minimizes the upper bound on the risk.

The work of Vapnik and Chervonenkis provided bounds on the risk functional for infinite sets of hypotheses such as the set of all hyperplanes in R^n. These bounds depend on the Vapnik-Chervonenkis dimension, which is the largest set of points, d, that can be correctly classified by a member of the set of hypotheses, H, for all the 2^d possible labellings (Vapnik, 1998). For example, the VC dimension of a linear decision function in two (n) dimensions is three $(n+1)$, since it is possible to correctly classify any labelling of three $(n + 1)$ points provided they are not arranged on a straight line (lower-dimensional affine subspace). The following bound illustrates the trade-off between minimising model complexity, measured by the VC dimension, and reducing the number of training errors.

Theorem 1 *For any probability distribution $\mathcal{D}$ on $X \times \{-1, 1\}$, with probability $1-\delta$ over l random examples S, any hypothesis $h \in H$ that makes k errors on the training*

set S has error no greater than

$$err(h) \leq \epsilon(l, H, \delta) = \frac{2k}{l} + \frac{4}{l}\left(d \log \frac{2el}{d} + \log \frac{4}{\delta}\right) \tag{1.2}$$

provided $d \leq l$.

where e is the base of the natural logarithm and d is the VC dimension.

This provides justification for reducing the number of training errors or empirical risk for a particular hypothesis class since the other terms are fixed by this choice (the training algorithm for multi-layer perceptrons, presented in Section 1.3, minimizes a measure of the empirical error). The alternative of finding a minimum of the above bound for nested hypotheses with increasing VC dimension and non-increasing empirical risk is termed structural risk minimisation. There are, however, several factors that can limit the above result's practical applicability. Firstly, the bound applies to the worst case and is therefore overly pessimistic for many distributions. For example, strong correlations between attributes may lead to the distribution being arranged on a sub-manifold of the input space with a lower VC dimension. The rationale for structural risk minimisation is that, even if the bound is not tight, it may still give an indication of the hypothesis with optimum generalization, although this is not guaranteed. It is also possible for a learning algorithm to achieve excellent generalization on a reasonably benign distribution even if the above bound exceeds unity (Cristianini and Shawe-Taylor, 2000).

Secondly, the bound may be difficult to calculate for some learning algorithms such as multi-layer perceptrons. And thirdly, the bound is not applicable to algorithms with infinite VC dimension such as the nearest-neighbour classifier[5]. However, it could be argued that the capacity of classifiers with infinite VC dimension

[5]The nn-classifier can produce a consistent hypothesis on any number of non-overlapping examples and therefore has infinite VC dimension.

is, in practice, limited by the size of the training set and that the bound suggests poor generalization. A margin-based approach constructs bounds on the generalization that also take into account the properties of the input distribution. These bounds suggest that, in some circumstances, the capacity of linear machines can be controlled, even in high-dimensional feature spaces (Vapnik, 1998).

1.2.2 Linear Learning Machines

Usually examples are in the form of a series of continuous or binary measurements that can be represented by an attribute vector in a Euclidean *input space.* The simplest hypotheses are linear decision functions, and these provide the basis for more advanced techniques such as multi-layer perceptrons and support vector machines. Binary classification is often represented by a real-valued function $f : X \subseteq R^n \rightarrow R$ with an input $\mathbf{x} = (x_1, \ldots, x_n)'$ assigned to the positive class if $f(\mathbf{x}) \geq 0$ and otherwise to the negative class. For linear learning machines $f(\mathbf{x})$ is a linear function of $\mathbf{x} \in X$ and can be written as

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \tag{1.3}$$

$$= \sum_{i=1}^{n} w_i x_i + b \tag{1.4}$$

where $(\mathbf{w}, b) \in R^n \times R$ are parameters that determine the function and $\langle \cdot \rangle$ denotes and inner product. This type of decision rule can be interpreted geometrically as a hyperplane in the attribute space defined by $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$ with a normal $\mathbf{w}$ and a geometric distance from the origin of $-b/||\mathbf{w}||$. Intuitively, the difficulty of a binary classification problem depends on the degree of separation between the two classes. An advantage of linear decision functions is that this 'separation' can be quantified using the geometric margin γ_i, which is the Euclidean distance of an example to

the separating hyperplane, provided that the weight vector is normalised such that $||\mathbf{w}|| = 1$.

$$\gamma_i = y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) \tag{1.5}$$

A $\gamma_i > 0$ indicates correct classification of $(\mathbf{x}_i, y_i)$. The margin of the training set S with respect to the hyperplane is

$$\gamma = \min_{i=1,\ldots,l} \gamma_i \tag{1.6}$$

The justification for support vector machines and other maximal-margin algorithms comes from a PAC result, which bounds the error of hyperplanes that linearly separate a dichotomy of classes in terms of the margin with a full definition (Cristianini and Shawe-Taylor, 2000) stated below.

Theorem 2 *Consider thresholding linear functions L with unit weight vectors on an inner product space X and fix $\gamma \in R^+$. For any probability distribution $\mathcal{D}$ on $X \times \{1, -1\}$ with support in a ball of radius R around the origin, with probability $1 - \delta$ over l random examples S, any hypothesis $f \in L$ that has margin $\geq \gamma$ on S has error no more than*

$$err(f) \leq \epsilon(l, L, \delta, \gamma) = \frac{2}{l} \left(\frac{64R^2}{\gamma^2} \log \frac{el\gamma}{8R^2} \log \frac{32l}{\gamma^2} + \log \frac{4}{\delta} \right) \tag{1.7}$$

provided $l > 2/\epsilon$ and $64R^2/\gamma^2 < l$.

This bound is a monotonically decreasing function of γ in the above range and suggests maximising the geometric margin. An important aspect of this bound is

that it does not depend on the dimension of the feature space[6]. This suggests that the increase in capacity of a learning machine in a high-dimensional feature space and the associated potential for overfitting can be controlled by ensuring that the classes are separated by a large margin.

The slightly surprising feature of the result is its non-invariance to translations or linear rescaling, as performing any linear transformation on the data and the corresponding decision function would generate an identical solution. The fact that this represents an upper bound means that the tightest error estimate is obtained when the data are linearly transformed to minimise equation 1.7. This complicates using the bound to estimate generalization but still provides justification for maximising the margin of two classes for a particular input distribution.

1.2.3 Maximal Margin Classifiers

The maximum margin classifier optimises the bound in Section 1.2.2 by attempting to separate the data with the optimal separating hyperplane. The scaling freedom in the definition of the hyperplane $(\mathbf{w}, b)$ is dealt with by including extra constraints on the functional margin of the two classes

$$y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 \qquad (1.8)$$
$$i = 1, \ldots, l$$

Minimising the 2-norm of the weight vector then corresponds to maximising the geometric margin for these *canonical* hyperplanes. The equation for the optimal separating hyperplane is given by the solution of the optimisation problem

[6] or the VC dimension, recall that $d = n + 1$ for linear decision functions.

$$\begin{aligned} &\text{minimise}_{\mathbf{w},b} && \langle \mathbf{w} \cdot \mathbf{w} \rangle, \\ &\text{subject to} && y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 \\ & && i = 1, \ldots, l, \end{aligned} \tag{1.9}$$

The Lagrangian is therefore

$$L_P(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle - \sum_{i=1}^{l} \alpha_i [y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1] \tag{1.10}$$

The above optimisation problem is quadratic with linear inequality constraints and is therefore convex (Cristianini and Shawe-Taylor, 2000). Convexity implies that any minimum must also be a global minimum. This means that the gradient of L_P with respect to $\mathbf{w}$ and b must vanish at the minimum

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} y_i \alpha_i \mathbf{x}_i = \mathbf{0} \tag{1.11}$$

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i = 0 \tag{1.12}$$

These equality constraints can be re-substituted into the primal Lagrangian to obtain the Wolfe dual form

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle \tag{1.13}$$

with the constraint $\sum_{i=1}^{l} y_i \alpha_i = 0$ enforcing the bias, b and $\alpha_i \geq 0$. *Maximisation* of the above objective function realises the maximal margin hyperplane. The solution of these types of quadratic programming problems is discussed in later sections.

The form of the solution for convex optimization problems can be visualised by considering a convex quadratic function of two variables $f(x, y)$. There exists one global extremum, which can be considered a minimum without loss of generality. Imposing the inequality constraint C, $y \geq x$ divides the space R^2 into two half-spaces with the half respecting the constraint termed the *feasible region.* If the global minimum lies inside the feasible region the solution is identical to unconstrained minimization $\alpha = 0$ and the constraint C is *inactive.* If the minimum lies outside the feasible region, the constrained minimum lies along the line $x = y$ and the *active* constraint acts as an equality, identically to the Lagrangian case. The Karush-Kuhn-Tucker complementarity condition summarises these two possibilities in the equation

$$\alpha(y - x) = 0 \tag{1.14}$$

The KKT condition also applies to the solution of the primal objective function and provides useful information about the form of the solution, viz:

$$\alpha_i[y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) - 1] = 0 \qquad i = 1, \ldots, l \tag{1.15}$$

This implies that only those inputs $\mathbf{x}_i$ with a functional margin of one are associated with a non-zero Lagrange multiplier α_i. These *support vectors* are points lying on the maximised geometric margin and are the only examples that contribute to the decision function

$$
\begin{aligned}
f(\mathbf{x}) &= \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \\
&= \sum_{i=1}^{l} y_i \alpha_i \langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b \\
&= \sum_{i \in sv} y_i \alpha_i \langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b
\end{aligned} \tag{1.16}
$$

Removal of all non-support vectors from training does not alter the solution of the objective function, provided there is a unique solution[7]. It should be noted that b does not appear in the dual problem but is easily calculated by using the KKT condition for any support vector.

The decision boundary for a linearly separable classification problem in two dimensions is shown as the solid line in Figure 1.3.

1.2.4 More Difficult Problems: Noise and Soft Margins

The maximal margin classifier is likely to produce excellent generalization on problems to which it may be applied. However, linear separation does not occur in many real problems, and this has led to several adaptations of the maximal margin classifier. If the input vectors are corrupted by noise, the classes may not be separable which would mean that equation 1.13 has an empty feasible region. Imposing linear separation is also undesirable for noisy problems with a narrow geometric margin defined by outliers. Slack variables are therefore introduced to allow violation of the margin constraint

[7]The form of any quadratic programming problem means that any local minimum must also be a global minimum, however, it is possible for degenerate solutions to exist. The degenerate solutions form simply connected regions of the parameter space. An analogous case for quadratic functions of two variables is $f(x,y) = (x-y)^2$ where degenerate global minima exist along the line $x = y$

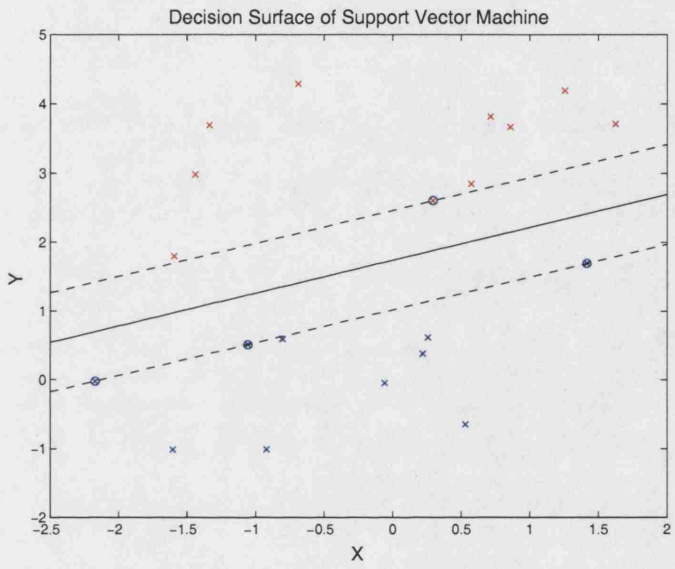

Figure 1.3: Decision surface of SVM for a linearly separable problem in two dimensions. The decision boundary $f(\mathbf{x}) = 0$ is shown by the solid line. The circled points are the support vectors, which lie on the dashed lines representing the geometric margin.

$$\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq +1 - \xi_i \quad \text{for } y_i = +1 \tag{1.17}$$

$$\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \leq -1 + \xi_i \quad \text{for } y_i = -1 \tag{1.18}$$

$$\xi_i \geq 0 \quad i = 1, \ldots, l \tag{1.19}$$

Each slack variable, ξ_i, represents the geometric distance to the margin hyperplanes for examples that fail to have a margin of γ with a slack variable that exceeds unity denoting a training error. An extra cost term must now be included in the objective function to penalize these margin errors, with an intuitively reasonable penalty being $\sum_{i=1}^{l} \xi_i$

$$\begin{aligned} \text{minimise}_{\mathbf{w},b} \quad & \langle \mathbf{w} \cdot \mathbf{w} \rangle + C \textstyle\sum_{i=1}^{l} \xi_i \\ \textit{subject to} \quad & y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \\ & \xi_i > 0, i = 1, \ldots, l, \end{aligned} \tag{1.20}$$

where C is a parameter that controls the trade-off between large margin and low empirical risk. The alternative 2-norm formulation places a quadratic cost on the slack variables and results in a quadratic programming problem that is identical to the maximal margin case except for a modification to the diagonal of the kernel matrix. Both 1 and 2-norm SVMs have been justified by PAC bounds on their generalization (Vapnik, 1998) and can be solved with similar efficiency. Equation 1.20 also illustrates the similarity of SVMs with regularisation networks with the second term penalizing empirical error and the first term acting as a quadratic weight decay regularizer (Bishop, 1995).

All margin errors (of which misclassifications are a subset) are included in the solution of the program and since many biological pattern recognition problems are

susceptible to outliers and have high empirical error, the 1-norm is used in most parts of this thesis. The primal Lagrangian for the 1-norm soft margin optimisation problem is

$$L_P(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \mu) = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle + C\sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i [y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) - 1 + \xi_i] - \sum_{i=1}^{l} \mu_i \xi_i \quad (1.21)$$

with the final term imposing $\xi_i \geq 0$. The dual problem is found by a similar argument to that used for the maximal margin and is identical except for the inclusion of an additional constraint (equation 1.23)

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle \quad (1.22)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C \quad (1.23)$$

$$\textstyle\sum_{i=1}^{l} y_i \alpha_i = 0 \quad (1.24)$$

This additional "box" constraint confines the vector $\boldsymbol{\alpha}$ to lie in the box of side length C in the positive orthant. Once again, the Karush-Kuhn-Tucker conditions provide information about the form of the solution

$$\begin{aligned} \alpha_i [y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) - 1 + \xi_i] &= 0, \qquad i = 1, \ldots, l \\ \xi_i(\alpha_i - C) &= 0 \qquad i = 1, \ldots, l \end{aligned} \quad (1.25)$$

The first constraint and the definition of the slack variable imply that points with a geometric margin greater than γ are associated with a null Lagrange multiplier and

do not contribute to the solution. All other points form the set of support vectors. The second constraint indicates that those support vectors with $\xi \neq 0$ i.e. margin errors, have a Lagrange multiplier at the upper bound ($\alpha_i = C$). A large fraction of these *bounded* support vectors is a characteristic of problems with a high empirical error. The support vectors that are not at the upper bound have a functional margin of 1 and lie on the geometric margin γ.

1.2.5 More Difficult Problems: Kernels and High-Dimensional Feature Spaces

The other limitation of linear learning machines with a VC dimension that is linearly dependent on the number of input attributes is that they may not have sufficient capacity to capture the *features* of the data that discriminate between two classes, or more succinctly, linear functions are not sufficient to separate the data with good generalization. In fact, linear learning machines are only useful for obtaining a multinomial combination of the class separations that are present in terms of the individual attributes (Bishop, 1995). This problem can be overcome by performing a non-linear mapping of the input vectors $\mathbf{x} \in X$ to a high-dimensional feature space F where a linear decision surface can separate the classes

$$\mathbf{x} = (x_1, \ldots, x_n) \mapsto \Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x})) \tag{1.26}$$

An important aspect of the dual representation is that input vectors only appear in the form of inner products. This means that explicit knowledge of the inputs is not necessary for solution of the dual problem once the Gram matrix $\mathbf{G} = (\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle)_{i,j}^{l}$ of the training set has been calculated. If a kernel function can be found that computes inner products in the feature space, F

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle \tag{1.27}$$

then it is not necessary to even know the precise form of the mapping $\Phi(\mathbf{x})$. The standard approach involves specifying a kernel function and then checking that this kernel corresponds to a valid inner product in some feature space. Mercer's theorem states the conditions for a function to admit a uniformly convergent expansion of the form

$$K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{z}) \tag{1.28}$$

where the λ_i can be viewed as weightings of the generalised inner product. Mercer's condition for a function to represent a valid kernel is that for all functions $g(\mathbf{x})$ with a finite L_2-norm

$$\int_X g(\mathbf{x})^2 \mathrm{d}\mathbf{x} < \infty \tag{1.29}$$

then

$$\int_{X \times X} K(\mathbf{x}, \mathbf{z}) g(\mathbf{x}) g(\mathbf{z}) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{z} \geq 0 \tag{1.30}$$

This condition also ensures that the kernel matrix of any training set will be positive semi-definite and that the dual problem will be convex. The simplest and most commonly used kernels are

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z} + 1)^p \tag{1.31}$$

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{z}||^2}{2\sigma^2}\right) \tag{1.32}$$

The first kernel results in a polynomial and the second a spherical Gaussian network with the maximal margin learning algorithm. Other non-Mercer kernel functions, such as the hyperbolic tangent

$$K(\mathbf{x}, \mathbf{z}) = \tanh(\kappa \mathbf{x} \cdot \mathbf{z} - \delta) \tag{1.33}$$

which results in a two-layer sigmoidal network can also be used, although a solution to the quasi-quadratic program may not exist.

An advantage of the SVM over radial basis function and multi-layer perceptron classifiers is that the number of hidden units is determined by the learning algorithm and does not need to be specified in advance. As a result, SVMs often achieve comparable performance to optimized RBF and MLP classifiers with relatively little tuning. Other kernels that also satisfy Mercer's condition can be generated by simple linear and multiple combinations of this initial set of kernels. Kernels can also be used to provide a geometric projection of non-Euclidean data into a geometric feature space. This is particularly useful for problems in biosequence analysis where examples are variable length strings of amino acids. An example is described in Jaakkola et al. (2000) where the parameters of a hidden Markov model are estimated for a particular protein family. The generative model is then used to estimate a similarity score between two query sequences, which then forms the kernel of a maximal margin classifier resembling the SVM.

1.2.6 Training Support Vector Machines

The main difficulty of solving the dual optimization problem becomes apparent when equation 1.13 is written in matrix notation

$$\text{maximize:} \quad L_D(\boldsymbol{\alpha}) = \boldsymbol{\alpha}'\mathbf{1} - \tfrac{1}{2}\boldsymbol{\alpha}'\mathbf{Q}\boldsymbol{\alpha} \tag{1.34}$$

$$\text{subject to:} \quad \boldsymbol{\alpha}'\mathbf{y} = 0 \tag{1.35}$$

$$\mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1} \tag{1.36}$$

where $\mathbf{1}$ and $\mathbf{0}$ are l-dimensional vectors of ones and zeros. The matrix Q is the Hessian $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$. Q is a symmetric matrix of size l^2 with $l(l+1)/2$ unique entries. This matrix cannot be stored for tasks with a large number of examples such as secondary structure prediction e.g. 100,000 examples corresponds to approximately 100Gb in float precision. Decomposition methods split the α_i into active (variable) and inactive (fixed) sets and optimise a series of smaller quadratic programs, which converge to the minimum of equation 1.34.

The elegant sequential minimal optimisation algorithm (Platt, 1999) computes analytical solutions for an active set of only two points, however, other implementations such as SVMlight (Joachims, 1999) can be trained considerably faster. The SVMlight package was used to train the SVM secondary structure classifiers in the following chapter. The loqo interior point optimiser was used as it is quicker than other quadratic program solvers for problems with a large number of bounded support vectors (Burges, 1998).

SVMlight uses a 'shrinking' heuristic to exploit the form of the solution with the majority of the Lagrange multipliers at the two bounds of equation 1.34 i.e. $\alpha_i = C$ for bounded support vectors and $\alpha_i = 0$ for non-support vectors. Lagrange multipliers that are found at either bound for a user-defined[8] number of iterations are fixed at the appropriate bound. The decomposition methods require that the training set fits into memory and the time complexity scales roughly with l^2 (the

[8]The default is 100.

time to compute $\mathbf{Q}$). Training time also scales linearly with the number of support vectors.

1.3 Feed-Forward Neural Networks

Although other learning algorithms such as SVMs and Bayesian networks have become popular in recent years, feed-forward neural networks trained using back-propagation continue to prevail as the standard approach for solving classification problems in bioinformatics. The chronological development of the back-propagation algorithm for training multi-layer perceptrons occurred in a similar order to the derivation of the SVM algorithm described in the previous section.

The natural starting point is the Widrow-Hoff procedure for obtaining a linear discriminant function for solving non-separable problems. This was then extended to non-linear discriminant functions by generalizing the Widrow-Hoff learning rule to networks with multiple layers of adaptive weights. As feed-forward neural networks are well established pattern recognition algorithms, this section presents only a brief sketch of their development with particular emphasis on the key differences with SVMs. More detailed discussion is given in Bishop (1995) and Duda et al. (2000).

1.3.1 Widrow-Hoff Learning Rule

The Widrow-Hoff or least-mean-square error learning rule was developed to obtain linear decision functions of identical form to Equation 1.3, except that the bias term is calculated as an extra weight, w_0, associated with a feature, x_0, which is set to unity for all examples. In matrix notation this gives

$$f(\mathbf{x}) = \mathbf{w}'\mathbf{x} \tag{1.37}$$

$$= \sum_{i=0}^{n} w_i x_i \tag{1.38}$$

where $\mathbf{x} = (x_0, x_1, \ldots, x_n)'$ is the original pattern vector augmented with the feature $x_0 = 1$.

The mean squared error $J(\mathbf{w})$ on a set of m training examples, written in matrix notation as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)'$, with labels $\mathbf{y} = (y_1, \ldots, y_m)'$ is therefore given by

$$J(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 \tag{1.39}$$

The weight vector, $\mathbf{w}$, that minimizes the objective function $J(\mathbf{w})$ can be calculated using the pseudoinverse of the rectangular matrix $\mathbf{X}$ (Duda et al., 2000; Golub and van Loan, 1996). However, the Widrow-Hoff learning rule avoids potential for loss of numerical precision or singular matrices occurring in the calculation of the pseudoinverse. The Widrow-Hoff learning rule involves minimising Equation 1.39 by gradient descent

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \eta(k)\nabla J \tag{1.40}$$

$$= \mathbf{w}(k) + \eta(k)2\mathbf{X}'(\mathbf{X}\mathbf{w}(k) - \mathbf{y}) \tag{1.41}$$

where k is the iteration of the optimization, which is initialized with a random weight vector, $\mathbf{w}(0)$. The learning rate parameter, $\eta(k)$, is reduced at each iteration to ensure convergence. Equation 1.40 suggests that the algorithm involves storage of the $n \times n$ matrix $\mathbf{X}'\mathbf{X}$. The space requirements can be reduced still further

by converting this *batch* algorithm to the *online* form where the weight vector is updated after presentation of each training example.

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \eta(k)(y^k - \mathbf{w}'(k)\mathbf{x}^k)\mathbf{x}^k \tag{1.42}$$

where y^k and $\mathbf{x}^k$ are the label and feature vector of the k^{th} training example to be presented to the network. This algorithm provides an efficient, general solution to solving linear problems that can be extended simply to non-linear problems, as described in the following section.

1.3.2 Multi-Layer Perceptrons

Non-linearity is incorporated into feed-forward networks by including a layer of 'hidden' processing units or neurons with a differentiable activation function and an additional layer of adaptive weights, as shown by Figure 1.4.

These networks are usually trained with variants of the error back-propagation algorithm, which again optimizes a mean-squared error function by gradient descent. The layer of adaptive weights connecting the hidden layer to the output node is trained using the Widrow-Hoff learning rule. The layer connecting the input nodes to the hidden layer is trained using the chain rule to obtain the gradient of the error function with respect to each weight. The error surfaces of multi-layer perceptrons are susceptible to containing local minima, which usually prevent gradient descent algorithms from reaching the global minimum. This presents relatively few problems in practice, as the global minimum does not necessarily provide better generalization than less optimal solutions. The standard approach to finding a solution is to initialize the network with a random weight setting and to optimise the weights until the algorithm converges or the error on a *validation* set begins to increase (early

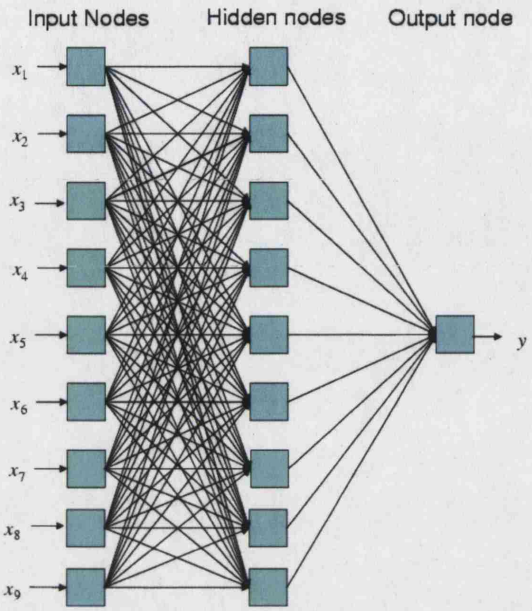

Figure 1.4: Feed-forward neural network. The dimensionality of the input vector, n, is 8 and there are also 8 hidden units.

stopping). This is then repeated several times with the most accurate network retained as the solution. Early stopping is one of several techniques, which include pruning and weight decay, that are designed to improve generalization by preventing the network from overfitting the training set (Bishop, 1995; Duda et al., 2000).

The capacity of neural networks is determined by several aspects of the network such as the topology and the magnitude of the adaptive weights. This provides a means for manually tuning the classifier to optimize performance, but ideally this tuning would be carried out automatically by the learning algorithm. One of the potential advantages of kernel-based techniques is that they may require less tuning and adjustment of arbitrary 'nuisance' parameters.

1.4 Structure of Thesis

One of the greatest unsolved problems in biology is in understanding how a sequence of amino acids drawn from a twenty-letter alphabet folds into a three-dimensional protein structure. The computational complexity of predicting the full 3D conformation of proteins has stimulated the development of knowledge-based approaches that solve simple intermediate problems such as predicting the secondary structure and local structural stability. These predictions are often core elements of methods that predict the structure of proteins that have negligible sequence identity to existing structures (Meiler and Baker, 2003). The first two chapters of this thesis describes some work on predicting the 1-dimensional properties of the protein structure. Although, secondary structure prediction has declined in importance in the period since the research in Chapter 2 was carried out, the chapter represents a useful introduction to some of the issues a researcher might encounter in applying SVMs to problems in bioinformatics. Other general aspects of machine learning such as fitting posterior probabilities to the outputs of a classifier, parameter optimization and benchmarking are also introduced in this chapter.

A similar approach is used in Chapter 3 to develop classifiers for identifying regions of the protein that do not adopt a stable conformation in their native state. This *native* or *intrinsic* disorder is defined as a lack of static, globular structure in the absence of a stabilizing interaction with nucleic acids, proteins or ligands. The problem of discriminating between ordered and disordered structures is used to determine which features characterize the two structural states, and to compare the SVM with feed-forward neural networks. The classifier DISOPRED2, which is shown to have a higher accuracy than other disorder prediction methods on the targets from the fifth Critical Assessment of techniques for protein Structure Prediction (CASP) experiment, is also used to investigate the frequency of disorder in complete

proteomes.

In Chapter 4, the DISOPRED2 classifier is used to provide greater insight into the biochemical functions of proteins that contain long regions of disorder in the eukaryote model organism *Saccharomyces cerevisiae*. This is carried out using the Gene Ontology annotations of the proteome supplied by the Saccharomyces Genome Database (Gene Ontology Consortium, 2000; Dwight et al., 2002). This chapter also includes a structural study of whether the linker regions between globular domains tend to be predicted as disordered. Although Chapter 4 deviates from the other three research chapters in not training classifiers using supervised learning, several relevant statistical techniques such as resampling (Efron and Tibshirani, 1993) and hypothesis testing (Weisstein., 2004) are used. The results from this chapter also motivate the approach to predicting protein function used in the final chapter.

Chapter 2

Predicting Protein Secondary Structure with Support Vector Machines

The large collaborative sequencing projects have yielded hundreds of bacterial and tens of eukaryotic genomes, and were instigated to give insight into the processes that create and maintain living cells in the three kingdoms of life. Amino acid sequences form the most important link between the information stored in DNA and the structure that carries out biological functions. Although the large structural genomics projects have enabled protein structures to be determined more quickly and efficiently than ever before, there remains a disparity between the rates of genome sequencing and experimental protein structure determination (see Figure 1.1). This has meant that little is known about the structure of many of the protein sequences that have been uncovered.

The gap between the rates of sequencing and experimental structure prediction has providing one stimulus for the development of computational approaches to predicting protein structure. The oldest and most established predictors of protein structure assign each residue in the sequence to the secondary structure classes of helix, strand or loop. Secondary structure prediction methods are often components of other algorithms for assigning the protein to a fold class or for calculating *ab initio* models of tertiary structure (McGuffin and Jones, 2002; Meiler and Baker, 2003).

Secondary structure prediction was amongst the first problems in biochemistry to be tackled computationally (Chou and Fasman, 1974) and has since been approached using numerous learning algorithms including multi-layer perceptrons and recurrent neural networks. At the time this research was carried out, secondary structure prediction represented a useful problem for testing the effectiveness of new techniques. Support vector machines (SVMs) had shown promising results on several biological pattern classification problems such as the recognition of protein translation-initiation sites in DNA sequences (Zien et al., 2000) and functional annotation of genes from expression profiles (Brown et al., 2000). Our initial study of protein secondary structure prediction was designed to investigate further the po-

tential costs and benefits of using SVMs for pattern recognition problems in bioinformatics. Several techniques and concepts introduced in this chapter are also used in subsequent chapters.

This chapter begins by giving a brief introduction to protein structure and the two regular secondary structure elements that allow close packing of the interior core of globular proteins. This is followed by a chronological review of developments in secondary structure prediction, up to and including the current state-of-the-art. The chapter also describes adaptations of SVM classifiers that provide confidence estimates for each prediction and that allow multiple-class predictions to be obtained from several binary classifiers. This is followed by the development of an SVM predictor and its comparison with several of the most accurate publicly-available modern methods. The chapter concludes with an evaluation of the advantages and disadvantages of using support vector machines for this problem, and discusses potential extensions to secondary structure prediction.

2.1 Predicting Protein Secondary Structure

Proteins perform a remarkably diverse range of functions within the cell; examples include recognition of pathogens, catalysis of metabolic reactions and regulation of gene expression. It has been shown in many experiments that the protein structure is vital for the proper functioning of the protein (Branden and Tooze, 1999), as breaking of the hydrogen bonds that stabilize the fold generally leads to deactivation of the protein. This requirement for a stable structure means that relatively few of the possible polypeptides of arbitrary length actually appear in nature (Rost, 2001a), as large areas of sequence space are not populated by stably folded proteins. However, there is growing interest in locally unstructured proteins, which are utilized in eukaryotic proteins for the reversible binding of DNA and the cytoskeleton (see

Chapter 3).

2.1.1 Protein Structure

The twenty types of amino acid that are present in most organisms form an alphabet with protein chains represented by a string of these residues[1]. Amino acids are composed of a carboxyl group, an amine group, a hydrogen atom and a side chain all bonded to a central α-carbon atom. The tetrahedral arrangement of four different groups, covalently bound to the chiral α-carbon, confers optical activity on amino acids[2]. Only the L-amino acids are translated from mRNA into protein although D-isomers are occasionally formed by post-translational modification. Polypeptide chains are created by the formation of covalent dipeptide bonds between the amide and carboxyl groups of two amino acids with the elimination of a water molecule. Proteins are formed by polymerization of these amino acid monomers at the ribosome where the sequence of nucleic acids in messenger RNA is translated into protein.

The peptide unit is rigid and planar, as resonance of the double carbon-oxygen bond restricts rotation. However, the single bonds on either side of the α-carbon allow rotation about the rigid peptide unit. The protein can therefore be thought of as a series of rigid links with free rotation about the covalent bonds on either side of the α-carbon. The rotations about the N–C_α and the C_α–C are denoted by the ϕ and ψ angles, respectively. Since these are the most important degrees of freedom, the conformation of the whole main chain of the protein is almost completely specified by these angles. A plot of the two torsional angles associated with the C_α atoms in a protein structure (named after the Indian biochemist Ramachandran) shows that

[1]The side-chains of certain amino acids can be altered by the addition of other chemical groups (e.g. phosphorylation of serine, threonine or tyrosine residues).

[2]The exception being the amino acid glycine, which has a side chain R composed of another single hydrogen atom.

the regular secondary structures are characterized by fairly distinct pairs of ϕ and ψ angles (Branden and Tooze, 1999). It is usually assumed that the protein structure is specified by the sequence of amino acids, although structures can be influenced by interactions with other macromolecules, metal ions, ligands and membranes.

The interior core of a globular protein consists of predominantly hydrophobic residues, which fold to minimise hydrophobic interactions with surrounding water molecules. The regular secondary structural elements α-helix and β-sheet allow the formation of hydrogen bonds between polar N–H and C=O groups in the main chain to entropically stabilize the protein structure. Van der Waals forces between hydrophobic side chains further stabilise the interior of the protein. These factors cause hydrophobicity to be the strongest determinant of protein structure, although other properties such as the size, geometry and chemistry of the side chain are also important. The following sections discuss the α-helix and β-sheet in more detail.

α-helix

α-helices are linear structures with the helix defined by the protein main chain and the side chains projecting outwards. The helix is stabilised longitudinally by the hydrogen bonds between the C=O group of residue n and the NH group of residue $n+4$ in the chain and is very stable compared with other structures. Some amino acids such as alanine and leucine have a propensity for forming α-helices and others such as valine and glycine oppose helix formation. The most well-known example, proline, is detrimental to helix formation because of steric hindrance by its rigid five-membered ring and the absence of an NH group for forming a hydrogen bond with the $n-4$ residue in the helix (Stryer, 1995).

The hydrophobicity pattern of an α-helix can indicate whether it is buried in the protein core or exposed on the surface of the protein. Buried helices are pre-

dominantly hydrophobic and helices at the surface typically have an alternating pattern of hydrophobic and hydrophilic residues (Branden and Tooze, 1999). Helices composed entirely of hydrophilic residues may have intermediate stability and be partially unstructured in solution, as discussed at greater length in the following chapter.

In the compact 3_{10}-helix and the loose π-helix, hydrogen bonds are formed between the C=O group of residue n and the NH groups of the $n+3$ and the $n+5$ residues respectively. These other types of helix are less favourable energetically and therefore rarer than the α-helix (see Table 2.1). They also tend to occur at the ends of α-helical structural elements.

β-sheet

The β-sheet differs markedly from the rod-like helical structures. The polypeptide chain of the strands in a β-sheet is almost fully extended with an axial separation of 3.5Å compared with 1.5Å for α-helix. β-sheets are formed when two or more parts of the chain are aligned so that hydrogen bonds are formed between the C=O of one strand and the adjacent NH groups of the other. Sheets are often described as being 'pleated' because the side chains project alternately above and below the plane of the sheet structure.

The chains can run in the same (parallel) direction from amino to carboxy ends of the protein or in opposite directions for anti-parallel strands. A single sheet structure can also be composed of a mixture of parallel and anti-parallel strands. Anti-parallel sheets are more common than the other two types because they can be formed by the chain folding back on itself to create a β-hairpin, whereas mixed and parallel sheets generally have a helical region in the intervening sequence. The isolated β-bridge has a similar bonding pattern to strands but is formed between

Type	H	G	I	A	P	E	B	C
Percentage	30.18	3.77	0.02	16.49	4.27	0.96	1.30	43.01

Table 2.1: Percentage of DSSP assignments (Kabsch and Sander, 1983) for a set of 5100 proteins with chains recorded in the protein data bank (Berman et al., 2000). The strand assignments from DSSP have been divided into anti-parallel strands (A), parallel strands (P), mixed strand (E). The other letters represent α-helix (H), 3_{10}-helix (G), π-helix (I), , β-bridge (B), and coil (C).

two isolated residues.

The two regular structures are connected by regions, which are typically static but have less spatial order. These regions are designated coil in the field of secondary structure prediction. The secondary structure elements tend to coalesce in the core of the protein with the coil regions exposed to the solvent. This results in a tendency for coil residues to be more hydrophilic and less evolutionarily conserved, since point mutations or insertions/deletions within coil segments are less likely to destabilize the protein structure and led to a loss of function. These conservation properties are one of the reasons for the improved prediction of secondary structure using evolutionary profiles rather than single sequences.

The overall secondary structure composition can be used as a very general classification scheme for protein structures. The classes are all-α, all-β, $\alpha + \beta$, where the helices and strands do not interact (strands are anti-parallel), and α/β structures, which are characterized by parallel strands and interacting structural elements (Murzin et al., 1995). There are two further organisational layers above the primary sequence and the secondary structure. The tertiary structure is the full conformation of the protein, as shown in Figure 2.1. Large sequences are often divided into distinct globular regions called domains separated by flexible segments. The quater-

Figure 2.1: X-ray structure of the tumor suppressor protein *p53* bound to DNA (red double helix). Helices are represented by rods, sheets by ribbons and coil structures by strings.

nary structure describes complexes comprising more than one protein chain or the structure of a single chain with multiple domains. Examples of quaternary structures include viral coat proteins and the tetramers that form active haemoglobin.

The structures of proteins that have high sequence similarity to existing structures can be determined fairly accurately by homology modelling algorithms. These methods exploit the fact that structure is highly conserved in related protein sequences (sequences that share more than 35% pairwise identity over 100 aligned residues adopt similar structures (Rost, 2001b)). For this reason, secondary structure prediction methods are only benchmarked on proteins that have low similarity with existing structures.

Some *ab initio* approaches to protein structure prediction attempt to use only the physical properties of the string of amino acids to fold the protein. Unfortunately, these methods are computationally intense and not particularly accurate. The best methods for predicting the structure of proteins that have little similarity to any existing structures use some physical principles in addition to databases of known structures (Jones, 2001; Rohl et al., 2004).

2.1.2 History of Secondary Structure Prediction

The regular structural elements, α-helices and β-sheets, were first postulated in the early 1950s by Pauling and Corey and were confirmed later in that decade by the first X-ray crystal structures of the protein myoglobin. However, the automatic approaches for predicting secondary structure did not arrive until a larger number of proteins were given some form of structural annotation in the mid 1970s. Nonetheless, secondary structure predictions were amongst the earliest examples of algorithms being applied to biochemistry. Much of this discussion is adapted from several reviews of the field by Burkhard Rost (Rost, 2001a; Rost and Sander, 2000; Rost, 1996), which organise its progress into three generations with significant conceptual developments marking the start of each generation.

The first computational method for determining these structures directly from sequence was developed by Chou and Fasman (1974), and used the single-residue preferences of each amino acid for forming the three structural states in addition to a set of simple rules to predict secondary structure. Secondary structures were defined using circular dichroism experiments, and the preferences P of an amino acid A for a particular structural state S were calculated by[3]

$$P_A(S) = \log\left(\frac{p(A,S)}{p(A)p(S)}\right) = \log\left(\frac{p(A|S)}{p(A)}\right) = \log\left(\frac{n_{A,S}/n}{n_A n_S/n^2}\right) \qquad (2.1)$$

where n_x is the number of residues in state x, giving the log odds of the conditional probabilities for each amino acid being associated with a particular structure. The rules essentially involve searching through the sequence for contiguous regions where $P(S = \text{helix}) > 0$ or $P(S = \text{sheet}) > 0$ and then extending these regions in both

[3]The original calculation by Chou and Fasman (1974) did not include the logarithm of the conditional probabilities but they are included in Equation 2.1 for consistency with score matrices such as BLOSUM62.

directions. This was followed by the GOR method (Garnier et al., 1978), which again used single-residue statistics that were averaged over a window of residues, and was trained using structures elucidated by X-ray crystallography. The improved definition of structures and the use of a window improved the prediction accuracy significantly over earlier methods.

The Dictionary of protein Secondary Structure program (DSSP), was another key development in structure prediction as it provided an objective means for assigning secondary structure from a set of atomic co-ordinates (Kabsch and Sander, 1983). Prior to DSSP, secondary structures were assigned manually, and therefore subjectively, by the crystallographer. DSSP uses length and geometry constraints to establish hydrogen bonds between adjacent C=O and N-H groups. These hydrogen bonding patterns are used to identify helix in addition to 'ladder' interactions, which form 'bridges' and then 'sheets' in the tertiary structure.

DSSP has persevered as the standard structural assignment algorithm for the past twenty years although the STRIDE method represents an attempt to improve upon it (Frishman and Argos, 1995). STRIDE also uses the torsion angles ϕ and ψ to assign structures, and is more accurate at assigning the ends of helical regions. Results from the CASP2 experiment indicate that the STRIDE and DSSP definitions differ for only 4.8% of residues although this can rise to 12% for some structures (Lesk, 1997). Both methods for assigning secondary structure are prone to error on low resolution structures because of a failure to recognize hydrogen bonds between poorly resolved atoms. Many modern methods avoid this difficulty by only training on structures with resolutions better than around 2Å, since this allows the crystallographer to remove most of the errors from the model. The DSSP assignments are more commonly used than STRIDE for benchmarking secondary structure prediction, and are used in later sections of this chapter. The second generation secondary structure prediction methods used windows of between 3 and 51

adjacent amino acids to predict the structural class of the central position. The most accurate of these was developed by Qian and Sejnowski (1988) and used cascaded feed-forward neural networks to predict three-state structure from binary-encoded amino acid sequence. This machine learning system is the antecedent of many of the later prediction methods including the current state-of-the-art (Jones, 1999).

The beginning of the third generation was marked by methods that make use of information from homologous sequences. Evolutionary information improves accuracies because the level of conservation and the properties of the substitute residues in related proteins also contain some implicit information on the longer range interactions and the environment. The first of the third-generation methods was PHD (Rost and Sander, 1993), which was the first to exceed 70% three-state prediction accuracy. PHD uses cascaded feed-forward neural networks that closely follow those used by Qian and Sejnowski (1988) on single sequences. The inputs are generated by performing a BLAST search on a sequence database to detect homologous proteins (Altschul et al., 1990). An $M \times 20$ position-specific scoring matrix (PSSM) is then obtained from a multiple alignment of the homologous proteins and the target, where M is the length of the target sequence.

The attribute vector for each example is constructed from a symmetric window containing the scores for thirteen adjacent residues along with more global measurements of the amino acid composition and length. These are fed into a two-layer, feed-forward neural network with an output node for each structural class. A window of thirteen of these outputs is then fed into a second neural network to filter the predictions and reinforce correlations in the class assignments between adjacent residues to remove, for example, helices that contain fewer than three residues. PHD also incorporates consensus predictions from several networks trained using balanced and unbalanced training sets.

2.1.3 Modern Methods

Since the publication of PHD in the early nineties, prediction accuracies have gradually increased to their current peak of between 76 and 78%. During this period, the greatest improvement was achieved by Jones' PSIPRED prediction method (5-6%), which uses PSSMs from Position-Specific Iterated BLAST (PSI-BLAST) searches (Altschul et al., 1997) as inputs to cascaded neural networks that resemble those of PHD. PSIPRED uses a network architecture with a very large number of hidden units and is trained using early stopping, but its main advantage appears to be the use of iterated BLAST searches, which recover more remote homologues of the target protein. Obtaining the scoring matrix in one step is also less expensive computationally than finding homologues using BLAST and then performing a multiple alignment. The other cause for the increase in prediction accuracy is the huge growth in the size of the sequence databases, which have led to a continual improvement in the quality of the PSSMs (Rost, 2001a).

The other methods with comparable performance to PSIPRED include SAM-T99sec (Karplus et al., 1998), which uses profile hidden Markov models to perform the initial search for homologous proteins, and PROFsec, which is an enhancement of PHD using PSI-BLAST profiles and other slight changes. Unfortunately, the precise details of both these techniques have yet to be published. Profile hidden Markov models are successful in recovering more remote homologues than BLAST. However, the inclusion of very remote homologues in the scoring matrix can lead to a reduction in accuracy as the structures become less conserved. Consequently, these methods do not appear to improve upon predictions made using PSI-BLAST profiles.

The major limitation of techniques based on a local windows is that the long-range interactions that particularly affect sheet formation are not explicitly encoded.

The SSpro structure prediction method attempted to address this difficulty with a bi-directional recurrent neural network (BRNN) that incorporates contextual information from both ends of the chain (Baldi et al., 1999).

The network used by SSpro is based on a first-order input-output hidden Markov model and is perhaps the most sophisticated learning algorithm to be applied to secondary structure prediction. However, Baldi et al. (1999) observed that the gradient of the error function tends to vanish with respect to inputs from distant positions, as the information from distant portions of the protein sequence is sparse and noisy. Amino acids that are in close proximity in the folded globular structure are likely to influence the secondary structure of the central residue but may not be situated nearby in the amino acid sequence.

The problem of the vanishing gradients was addressed by using input windows from adjacent positions in the chain but leads to a prediction that is based on an effective window length of around 31 amino acids. Consequently, SSpro does not appear to predict sheets any more successfully than the multi-layer perceptrons based on a narrower input window. SSpro also uses ensemble averages from several networks with varying numbers of hidden units, window lengths and forward-backward iterations. The latest version of SSpro uses profiles and also classifies structure into eight rather than the standard three states (Pollastri et al., 2002). The three-state prediction accuracy is almost identical to PSIPRED although the two methods appear to make different errors (McGuffin and Jones, 2002).

The study, described here, is not the first to apply support vector machines to the secondary structure prediction problem. Hua and Sun (2001) used multiple sequence alignments from the Cuff and Barton data set of 513 non-redundant proteins (Cuff and Barton, 1999) to train several binary SVMs with radial basis function kernels. Their best quoted Q_3 score of 73.5% is lower than that found in the present study

although there is substantial variation between test sets and a direct comparison cannot be made. Their prediction method was, however, benchmarked against PHD and the meagre improvement suggests that it is not competitive with the latest methods. A surprising feature of Sun and Hua's results is that the segment overlap score (Zemla et al., 1999) of 76.2% exceeds the three-state accuracy, despite the absence of a filtering network used by PHD and PSIPRED. A definition of the segment overlap score is given in Appendix B. The present work is inconsistent with their results as it suggests that, even with a significantly higher three-state accuracy, a second classifier is necessary to obtain similar segment overlap scores.

The following section describes the development of a new method for the prediction of secondary structure using SVMs.

2.2 Adapting SVMs for Secondary Structure Prediction

The standard SVM is essentially a binary classifier which, when given a test example, outputs the geometric distance to the optimal separating hyperplane. This presents certain difficulties in secondary structure prediction which involves multiple classes and where some form of confidence estimate is necessary to allow the user to interpret the quality of the prediction at different locations in the sequence. The following two sections describe approaches for combining binary classifiers and to obtaining posterior probability estimates.

2.2.1 Probabilistic Outputs

The three-state SVM prediction method combines results from several binary SVMs, so probability estimates are useful for making an overall prediction. The posterior probabilities are also necessary for optimising more general loss functions (Duda

et al., 2000). For example, the highest three-state accuracy may not be optimal for fold recognition methods that incorporate secondary structure, since under-predictions (where helix or sheet elements are classified as coil) appear to be more detrimental than over-predictions (McGuffin and Jones, 2002). The confidence estimates developed by Platt (2000) were used in this work since they can be implemented as a simple and fast to evaluate 'wrapper' around the standard binary SVM. The raw outputs of the SVM are mapped to posterior probabilities using a logistic sigmoid function

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)} \tag{2.2}$$

where $f(\mathbf{x})$ is the output of the SVM and the parameters A and B are found by maximum likelihood estimation. This choice is heuristic, and it is possible that the posterior probability follows a different distribution (Tipping, 2001). However, the approximation is found to be accurate in later sections.

The training set for the estimation is determined by three-fold cross-validation rather than classifying the training set because of the discontinuity at the margin. This causes the training set to be a biased sample because of the unbounded support vectors with a functional output of exactly 1, a value that is unlikely to occur frequently in testing. The training set is divided randomly into three with an SVM trained on two out of the three subsets and the outputs evaluated on the remaining third. The three sets of outputs are then combined and used to estimate the parameters of the model. A non-zero value for the parameter B indicates that the sigmoid has a decision threshold that differs from the raw SVM.

2.2.2 Multi-Class Classification using SVMs

One of the outstanding problems with support vector machines is applying a binary classification method to problems involving multiple classes. Several multiple-class SVMs have been discussed in the literature (Weston and Watkins, 1998; Crammer and Singer, 2001) but these are impractical for large problems (Hsu and Lin, 2002) because of the increased time complexity and it is necessary to resort to combining results from several binary classifiers.

The 'one-versus-rest' method constructs a binary classifier for each of the k classes. The ith SVM is trained on all l examples with those in class i given positive labels and examples belonging to all other classes given a negative labelling. The final decision for the 'one-versus-rest' classifier is the class corresponding to the SVM with the highest output value

$$\text{class of } \mathbf{x} = \max_{i=1,\ldots,k} f(\mathbf{x}, \boldsymbol{\alpha}_i, b_i) \tag{2.3}$$

This is justified on the grounds that the posterior probability is monotonic with respect to the functional output, as suggested by Equation 2.2, although there is a requirement that the scaling parameters A_i are similar for all classifiers. This assumption can be removed by calculating probabilistic outputs and assigning example $\mathbf{x}$ to the class with highest posterior probability. Unfortunately, the training and test times scale linearly with the number of classes for this method.

The 'one-versus-one' and other related methods construct classifiers for the $\binom{k}{2} = k(k-1)/2$ possible pairings with each classifier trained on the subset of the examples belonging to the two classes. In testing, the outputs are combined by casting a vote for the 'winner' of each pair-wise comparison and assigning the example to the class with the most votes. A disadvantage of the 'one-versus-one' classifier is the high

occurrence of tied results for problems with a small number of classes. These ties must be settled by some heuristic such as selecting the class with highest prior probability.

An alternative approach is pair-wise coupling (Hastie and Tibshirani, 1998), which combines probabilistic outputs from the binary classifiers (previous section) to obtain an overall prediction. The outputs are placed in an observation matrix $\mathbf{R}$ containing the outputs of the coil/helix r_{12}, coil/sheet r_{13} and helix/sheet r_{23} classifiers, an example being

$$\begin{pmatrix} \cdot & 0.77 & 0.65 \\ 0.23 & \cdot & 0.19 \\ 0.35 & 0.81 & \cdot \end{pmatrix}$$

The outputs are modelled by

$$P(\text{class } i | \text{class } i \text{ or class } j) = \mu_{ij} = \frac{p_i}{p_i + p_j} \tag{2.4}$$

which represents a system with $k(k-1)/2$ equations and $k-1$ free parameters since $\sum_i p_i = 1$. A solution that satisfies all of the constraints does not generally exist and the problem is overdetermined. The multi-class probabilities are therefore obtained by minimising the weighted Kullback-Leibler divergence between the model and outputs

$$l(p_i, \ldots, p_k) = \sum_{i<j} n_{ij} \left(r_{ij} \log \frac{r_{ij}}{\mu_{ij}} + (1 - r_{ij}) \log \frac{1 - r_{ij}}{1 - \mu_{ij}} \right) \tag{2.5}$$

where the n_{ij} are weights to account for the different precisions in the pair-wise probability estimates. These weights are typically the number of examples used to

train each classifier but they appear to have little effect on the overall predictions. A simple gradient descent algorithm is used to estimate the probability vector $\mathbf{p} = (p_i, \ldots, p_k)$ which is $\mathbf{p} = (0.53, 0.11, 0.36)$ for the example above. The algorithm typically converges in a few hundred iterations for three classes and does not add greatly to the time taken to classify an example.

Another method is the decision directed acyclic graph (DAG), which is trained in the same way as the 'one-versus-one' classifier but is justified by a PAC bound on its generalization (Platt et al., 2000). A rooted graph is constructed with k leaves and a binary SVM at each of $k(k-1)/2$ nodes. A test sample is classified by starting at the root node and evaluating the binary decision function. A positive output leads to an exit via the right edge of the node and a negative output to exit via the left edge. The example is then evaluated by a further $k-2$ binary SVMs in its path to a leaf node that indicates the predicted class. The training times for the 'one-versus-one' and related classifiers depend on several factors such as proportion of examples in each class and the number of support vectors.

The results of the various methods for combining binary classifiers into an overall prediction are given in the following section.

2.2.3 Data Preparation and Attribute Selection

A large set of proteins with solved crystal structures was clustered such that no two proteins had $> 25\%$ sequence identity between clusters. A representative structure with the best resolution was then selected from each cluster and placed in the training set, totalling 1460 non-redundant proteins. A position-specific scoring matrix was calculated for each of these sequences using three iterations of a PSI-BLAST search (Altschul et al., 1997) against a sequence database that was filtered to remove low sequence complexity, coiled-coils and transmembrane helices. The scores at each of

the M positions represent the log-likelihoods for each of the twenty amino acids at that site, calculated from the sequences recovered by BLAST.

The inputs to the SVM were constructed by considering a window of fifteen residues with the twenty scores for each residue forming a 300-dimensional attribute vector (see Section 2.2.4). At the ends of the protein where the window extends beyond the terminal residues, the profiles for the missing residues were assigned values of zero. This approach differs slightly from PSIPRED (Jones, 1999) and others (Hua and Sun, 2001; Rost and Sander, 1993) since the inclusion of extra inputs to indicate the ends did not alter the performance of the SVM and were therefore omitted.

Most prediction methods reduce the eight states outputted by DSSP to the helix (H), strand (E) and the default coil (C) states. One of the reduction schemes converts α-helix and 3_{10}-helix to H, sheet and isolated β-bridge to E and the rest (π-helix, turn, bend and other) to coil. The other converts only α-helix to H and strand to E with the other states becoming coil. The second scheme generally gives higher prediction accuracies because the 3_{10}-helix and β-bridge states are more difficult to distinguish from coil than either α-helix or strand. The first reduction scheme has now been accepted as a standard and is used in the rest of this chapter to allow comparison with PSIPRED (Jones, 1999) and other prediction methods. The scheme does, however, differ very slightly from the EVA[4] definitions (Rost and Eyrich, 2001), which also convert I to H but the scarcity of π-helices (Table 2.1) means that this does not greatly affect training or the testing accuracy.

[4]The EVA server maintains a continuous evaluation of various structure prediction methods, including secondary structure (Rost and Eyrich, 2001).

2.2.4 Effects of Window Length on Prediction Accuracy

An investigation of the effects of window size on the accuracy of the binary classifiers was carried out on symmetric windows with lengths varying from 5 to 23 residues. The general trend is that the performance of the SVM increases as the length of the window increases up to a certain threshold with a subsequent slow decline (Figure 2.2). The shallow decline illustrates the SVM's robustness to the addition of noisy and potentially redundant attributes as the window size increases. A window size of 15 was found to be roughly optimum for all three classifiers.

The primal weights of the linear classifiers with window length 15 are shown in Figures 2.3 and 2.4. These are calculated from the standard dual form of the SVM solution (see section 3.5) and show explicitly the parameters of the decision function, which is linear with respect to the input space. The relative sizes of the weights associated with each input attribute can therefore be used to indicate whether a high score for a particular feature contributes towards a positive or negative prediction.

The pattern shared by the helix and sheet weight plots in Figure 2.3 is a periodicity with 'wavelength' twenty, modulated by a function that peaks at the central residue. The distribution for the helix class is asymmetric and indicates that the 'downstream' residues have greater influence on helix formation. The large, positive, downstream weights are those associated with alanine, glutamine and leucine, with negative weights for proline and valine. This indicates that the helix-forming or -breaking properties of these downstream residues is more important for predicting whether the central residue forms part of a helix than the properties of the 'upstream' residues.

The weights for the central residues (Figure 2.4) also confirm some of the propensities for helix formation with alanine, glutamic acid and leucine tending to form

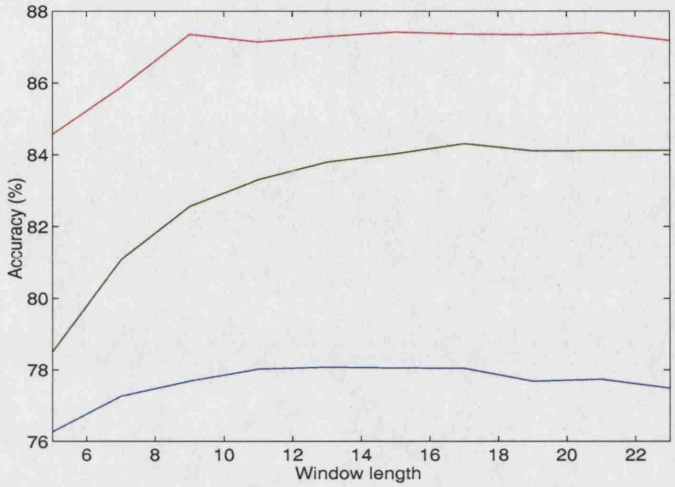

Figure 2.2: Hold-out accuracy of linear coil/¬coil (blue), sheet/¬sheet (green) and helix/¬helix classifiers with varying window length.

helices, and proline, glycine and serine opposing helix formation. Proline also deters sheet formation because it cannot participate in the hydrogen bonding that stabilizes the regular structural elements. Glycine's small side chain allows it to adopt Ramachandran angles that are outside those that characterize helices and sheets (Branden and Tooze, 1999).

The weights for the coil classifier have less periodic character and are more peaked at the central residue. Interestingly, two of the largest weights are outside the central position and are associated with the scores for cysteine, suggesting that disulphide bonds in the adjoining C-termini positions prevent the formation of regular structures. The sheet classifier appears to have greater dependence on positions at the ends of the window than either coil or helix which again suggests that strand formation is governed by longer range interactions. The aliphatic residues valine and, to a lesser extent, isoleucine have propensities toward sheet formation.

The linear coil and sheet SVMs were fairly successful in predicting secondary

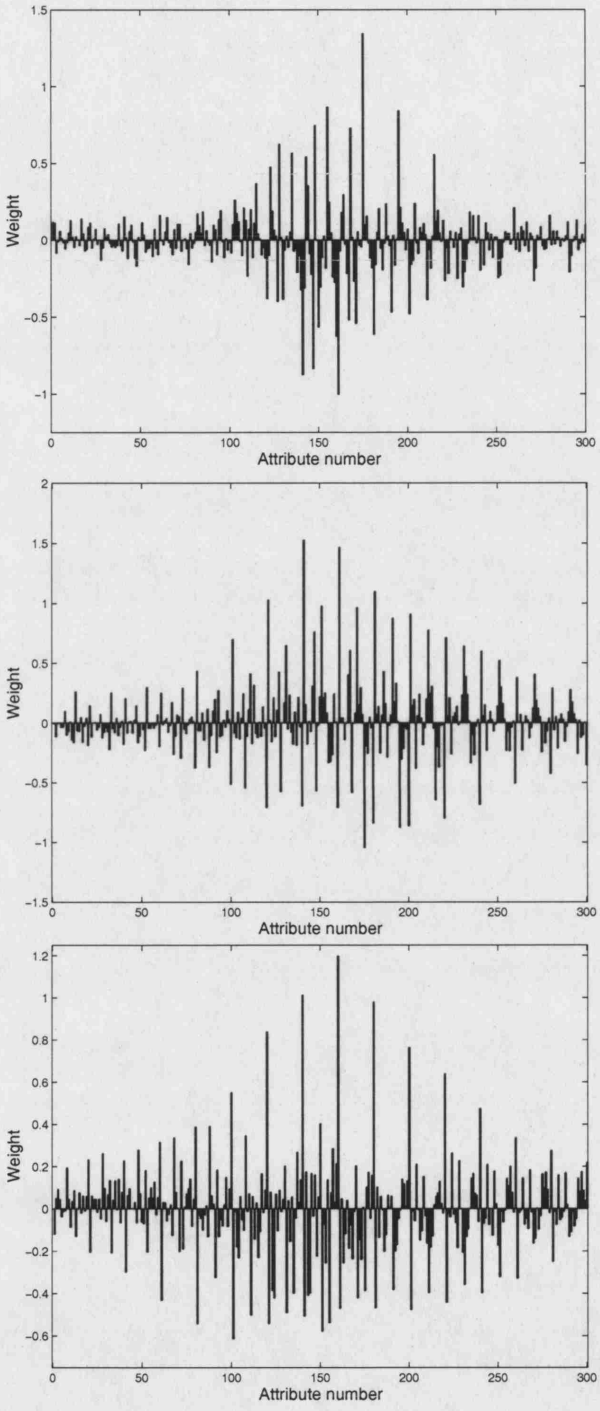

Figure 2.3: Plots in order from top to bottom: weights of linear SVMs for coil/¬coil, helix/¬helix and sheet/¬sheet classification. The first twenty attributes are the profile for the residue at the N-terminal end of the input window, twenty-one to forty the profile for the next residue in the window etc.

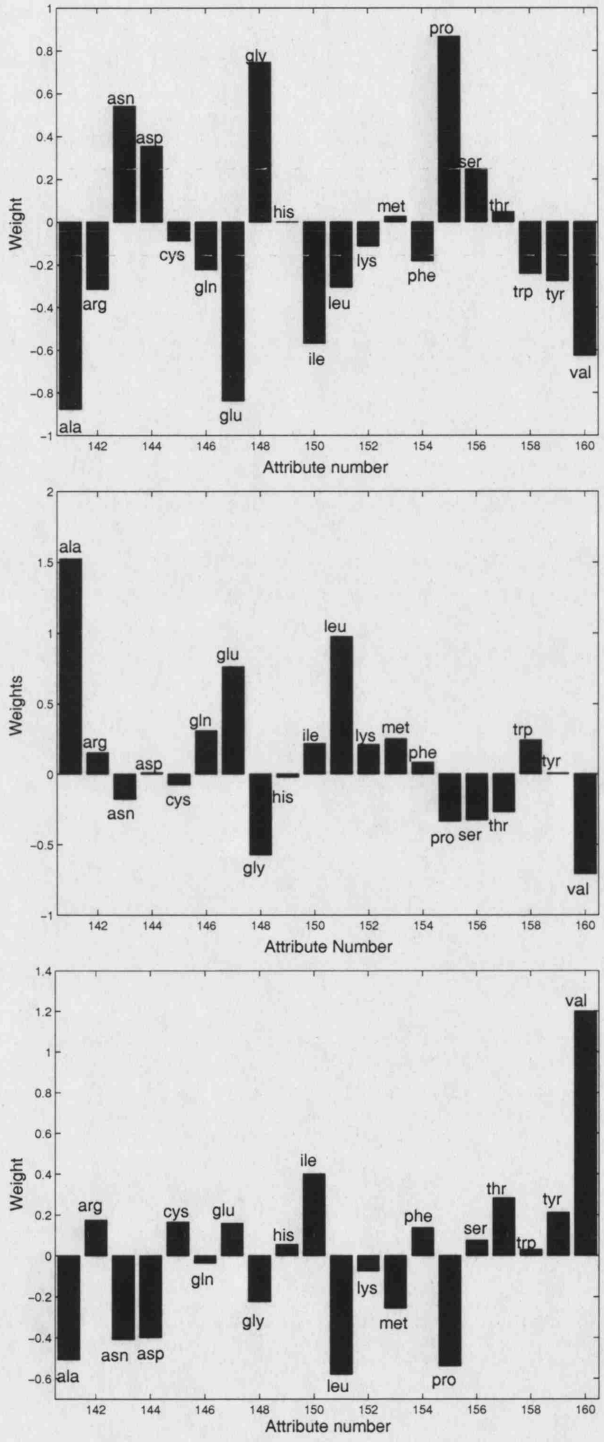

Figure 2.4: Weights of linear coil/¬coil, helix/¬helix and sheet/¬sheet SVMs for the profile of the central residue (attributes 141-160 of Figure 2.3).

structure, with an accuracy only a few percentage points lower than the best-performing non-linear kernel. However, the use of such a kernel gave more substantial improvement for helix prediction. This may be because non-linear SVMs have features that represent interactions between attributes and can therefore recognise periodic patterns in the hydrophobicity that signify a helical region (See Equation 2.7).

2.3 Results

The kernel function was found by optimising performance on a validation set. The training and validation sets were selected randomly from the filtered set of 1460 proteins, and contained 300 proteins (72583 examples) and 100 proteins (23785 examples), respectively.

2.3.1 Determining Kernel Parameters and Combining Binary Classifiers

Polynomial kernels were found to outperform the linear, Gaussian and sigmoid kernels. The best accuracy was achieved using second and third order polynomials. The quadratic kernel was chosen because it has faster training and classification rates than higher order functions. The following kernel function was therefore used

$$K(\mathbf{x}, \mathbf{z}) = \left(\frac{\langle \mathbf{x} \cdot \mathbf{z} \rangle + 1}{50} \right)^2 \tag{2.6}$$

where $\mathbf{x}$ and $\mathbf{z}$ are two example vectors with the scaling factor chosen to ensure that $K(\mathbf{x}, \mathbf{z})$ is in the range $[-1, 1]$. The optimal regularisation parameter was found by varying C through a range of values and evaluating the SVM's performance on the

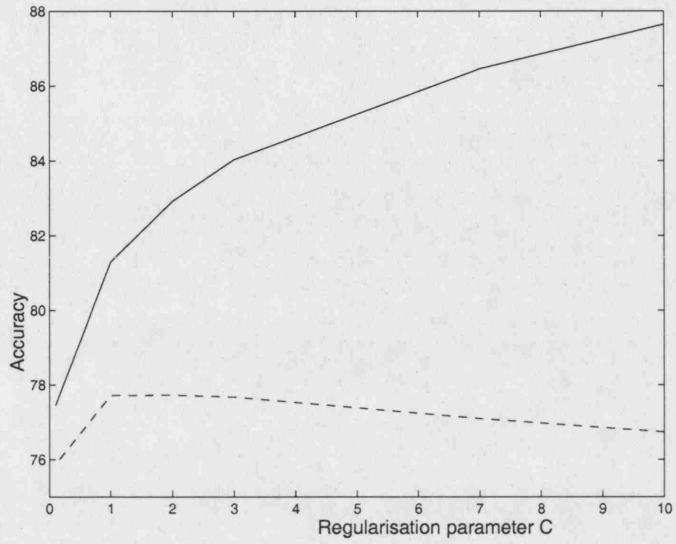

Figure 2.5: Dependence of training (solid curve) and validation set accuracy (dotted curve) on the regularization parameter C for the coil/ ¬coil classification problem.

validation set (Figure 2.5).

The SVM attained best performance on the validation set for $C = 2$ although the variation from $C = 1$ to $C = 3$ is not statistically significant. The decision function arising from a support vector machine, trained with the above kernel, can be viewed as a non-linear hypersurface in the input space or an affine function in the feature space consisting of the $\binom{d+1}{2}$ monomials of degree 2, the d input attributes and a constant (Burges, 1998); as shown by the equation

$$(\langle \mathbf{x} \cdot \mathbf{z} \rangle + s)^2 = \left(\sum_{i=1}^{d} x_i z_i + s \right) \left(\sum_{j=1}^{d} x_j z_j + s \right) \tag{2.7}$$

$$= \left(\sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j z_i z_j + 2s \sum_{i=1}^{d} x_i z_i + s^2 \right) \tag{2.8}$$

where is s is a scale factor that controls the trade-off between linear classification

Classifier Classifier	Support Vectors at upper bound (%)	Accuracy (%)	Above random (%)
C/¬C	55.0 (48.8)	77.7	20.5
H/¬H	40.9 (34.9)	86.4	19.7
E/¬E	36.5 (30.4)	85.6	9.5
C/(C ∨ H)	46.1 (39.5)	84.2	27.9
C/(C ∨ E)	48.5 (40.7)	81.3	17.1
H/(H ∨ E)	36.0 (29.6)	88.0	29.8

Table 2.2: The percentage of the examples in the training set that form support vectors and accuracy on the test set. The final column shows the SVM's improvement over the trivial prediction (class with highest prior probability). The training set contains coil=40.8%, helix=35.8% and sheet=23.4% residues.

and the non-linear terms.

Thus, high scores for alanine in the central residue and glutamic acid at the $n+4$ position, which might be expected to indicate a helical segment, can be represented by the SVM. The parameters of the function that maps the outputs of the SVM to posterior probabilities were estimated using the three-fold cross-validation with the procedure given in section 2.2.1. Instances where the sigmoid mapping is a poor approximation to the posterior probability have been shown (Tipping, 2001). The posterior appears to be roughly sigmoidal for secondary structure prediction but the intermediate estimates $0.1 \leq P \leq 0.9$ are biased by misclassified examples with large functional outputs. A better fit can be obtained by re-estimating the parameters on only those examples with outputs close to zero (Figure 2.6).

Six polynomial binary classifiers were trained on the set of 300 proteins to compare the three possible methods for predicting secondary structure (see Table 2.2).

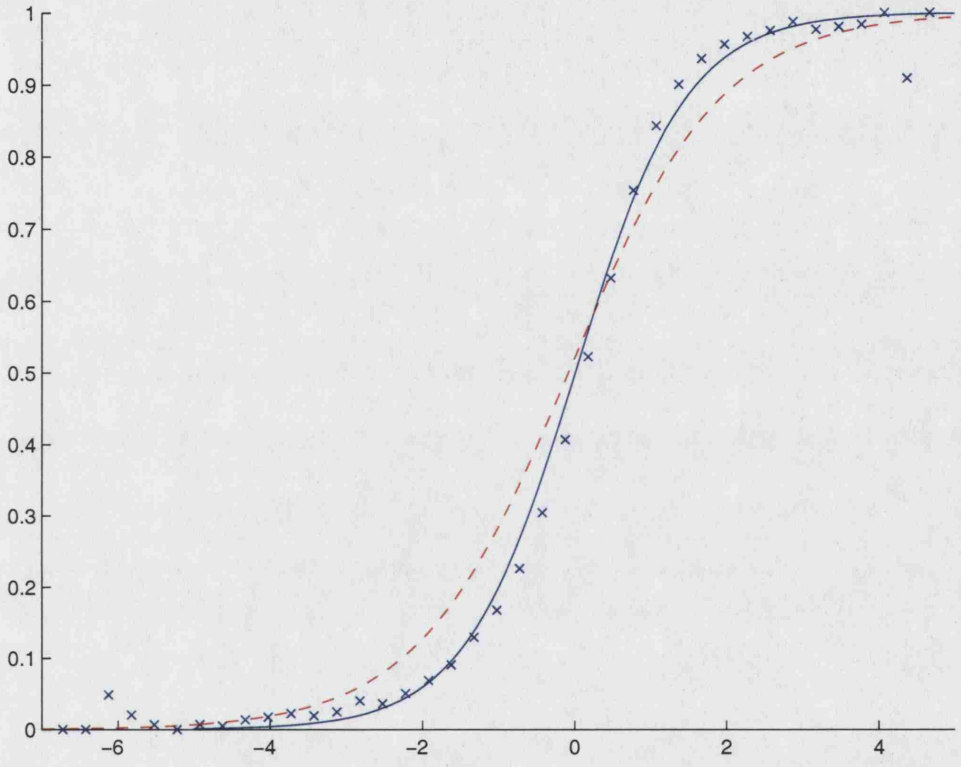

Figure 2.6: Logistic sigmoid fitted to the outputs of the coil/helix classifier using the method detailed in Platt (2000). The x-axis shows the functional output $f(\mathbf{x})$ and the y-axis the posterior probability $p(\text{Coil}|\text{Coil or Helix})$. The dashed red curve shows a sigmoid fitted to all examples, whilst the solid blue curve shows the fit to examples with $|f(\mathbf{x})| < 2.5$. The crosses show the posterior probability for points falling in bins of width 0.3.

The accuracy of 'one-versus-rest' classifiers suggest that coil and helix states are more easily discriminated than sheet. The pair-wise comparisons involving helix are also more accurate than coil/sheet classification. This is probably a result of helix formation being dominated by short-range interactions, which can be represented by the window of fifteen residues. The survey of protein structures in section 2.1 also indicates that some amino acids have strong propensities for either forming or breaking helices and that helical structures have clear patterns of hydrophobicity. Strands, on the other hand, are formed between two complementary strings of residues that are often distantly separated in the protein sequence. These longer-range interactions are not fully encoded by local windows, although the profiles do indicate residues that are under evolutionary constraint and therefore likely to form secondary structure elements.

The number of support vectors is extremely high for all six binary classifiers. This is undesirable because the fraction of the training examples that become support vectors forms a loose upper bound on the generalisation error rate of the support vector machine (Cristianini and Shawe-Taylor, 2000). The time complexity, in testing, also scales linearly with the number of support vectors (see Equation 1.16). The high empirical error is a characteristic of secondary structure prediction and is caused by noise in the evolutionary profiles and some ambiguity in the structure assignments. This high error leads to a large fraction of the data set becoming bounded support vectors.

The examples are also not independently and identically distributed (i.i.d), with the feature vector for a particular residue closely resembling a reflection about the origin of the feature vectors of adjacent positions in the protein chain. Results for the *virtual* SVM method[5] (Burges and Schölkopf, 1997) indicate that correlations

[5]Burges and Schölkopf (1997) trained a standard binary SVM to recognise isolated handwritten digits. Domain knowledge was then incorporated by applying the known size invariance to the

led to an increase in the number of support vectors. The number of support vectors is also known to grow linearly with the size of the training set.

The binary classifiers, shown in Table 2.2, were combined using several methods to obtain a full three-state prediction. The results of these combinations are shown in Table 2.3 along with results from a neural network. The neural net has a similar architecture to PSIPRED with a single hidden layer of 65 units and three output nodes. This network is trained using batch resilient back-propagation (Riedmiller and Braun, 1993) with early stopping on another 100-protein validation set. The scores are quoted for the 75 proteins without unresolved breaks from the test set of solved crystal structure.

In Table 2.3, the scores for the 'one-versus-rest', maximum probability and pairwise couple methods are significantly different to the DAG SVM and neural network (at the 95% level) according to a paired t-test. The differences between the 'one-versus-one' classifier and the neural network are all statistically significant apart from SSEA. The score quoted for the DAG SVM has the coil/helix classifier at the root. The other two arrangements had Q_3 scores within 0.13% of this upper value.

Table 2.3 suggests that, on a restricted data set, the various types of multi-class SVM predictors attain higher prediction accuracy than a comparable neural network. Although it is possible that the neural network could be improved by further tuning, the topology and learning algorithms are almost identical to PSIPRED, which has been optimized for the secondary structure prediction problem.

The segment overlap scores (Sov) (Zemla et al., 1999) and the secondary structure element alignment (SSEA) scores (McGuffin and Jones, 2002) for all six SVM

original set of support vectors to generate *virtual support vectors*. The SVM retrained on the virtual SVs had better generalisation but at the expense of approximately a factor of two more support vectors.

Classifier	Q_3	Sov	SSEA	C_C	C_H	C_E	Wrong
one-versus-rest SVM	74.54	68.24	73.04	0.55	0.68	0.59	3.0
maxprob SVM	74.52	68.45	73.07	0.55	0.68	0.59	2.9
one-versus-one SVM	74.04	67.11	72.16	0.54	0.68	0.58	2.5
DAG SVM	74.14	66.95	72.19	0.54	0.68	0.58	2.6
pair-wise couple SVM	74.24	68.34	72.94	0.54	0.68	0.59	2.7
neural network	73.28	64.65	72.19	0.54	0.66	0.58	3.6

Table 2.3: Q_3, Sov (Zemla et al., 1999) and SSEA (McGuffin and Jones, 2002) scores for binary classifiers on a test set of 75 proteins and comparison with a neural network (Q_3 Sov and SSEA scores are defined in the Appendix). C_X are the Matthew's correlation coefficients for coil, helix and strand. The wrong score represents the percentage of residues where helix is misclassified as sheet and vice versa.

classifiers are significantly lower than most modern prediction methods. The following section discusses the reasons for the low Sov and SSEA scores and describes the development of an array of structure-to-structure classifiers for improving Sov, SSEA, and to a lesser extent, Q_3 scores.

2.3.2 Training a Set of Structure-to-Structure Classifiers

The Sov score was designed to be a more realistic measure of the quality of secondary structure predictions by penalizing misclassification of whole structural elements more severely than errors in length. It is therefore less affected than Q_3 by ambiguities at the ends of structural elements that may arise in the actual 3D structure. SSEA is calculated by performing a global alignment of the secondary structure elements in the prediction and the target structure, and gives different weights to the various types of error (see Appendix). The predictions in Table 2.3 have relatively low Sov and SSEA scores because of occasional inconsistent assignments within a

layer 1: HHCECEHECHHHCC

layer 2: HHHHEEEEEEHHHCC

Figure 2.7: Example predictions for protein 1qkr(A) from the EVA (Rost and Eyrich, 2001) server. The middle segment, which contains all three structure types in the first-layer prediction, has been converted to a continuous strand element by the second layer.

single structural element as shown in Figure 2.7.

This work follows others in the secondary structure prediction field (Rost and Sander, 1993; Jones, 1999) and trains a second classifier to filter the predictions. An SVM trained directly on the outputs obtained from the training set does not improve greatly on the first set of predictions, which is probably another consequence of the discontinuity at the margin. However, the three-state class probabilities can be used to train a second structure-to-structure SVM. The accuracy of the second set of SVMs did not appear to depend on whether the 'one-versus-one' or the 'one-versus-rest' inputs were used, so the pair-wise coupled inputs were used as each test example is evaluated on a factor of 1.8 fewer support vectors.

The probability of coil is simply a linear function of the helix and sheet probabilities and was not included in the inputs to the second set of SVMs, which were constructed from another window of fifteen residues. The probabilities of coil and helix were rescaled by subtracting 1/3 so that null vectors, used for the positions beyond the protein termini, represent an equal probability for the three-structure types. The inclusion of an extra attribute to signify the end positions was also found to improve the accuracy.

The spherical Gaussian kernel outperformed both the linear and higher order polynomials for the second set of predictions. The parameters C=0.2 and σ=0.625 were found using the *looms* (leave-one-out model selection) program (Lee and Lin,

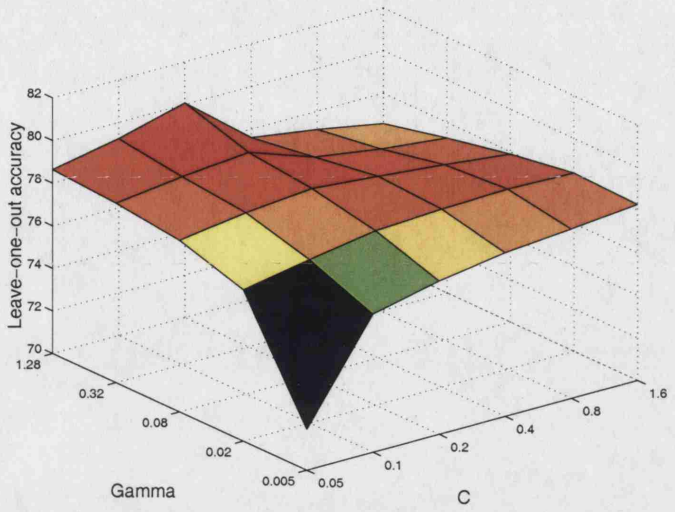

Figure 2.8: Leave-one-out estimates of prediction accuracy for coil/¬coil classifier on training set of 300 proteins.

2000), which uses loose stopping criteria to calculate estimates of the leave-one-out error for a range of values of C and $\gamma = 1/2\sigma^2$ as shown in Figure 2.8. The looms program does not provide accurate estimates of the generalisation but, in practice, the classifier with the highest prediction accuracy is expected to have optimal kernel parameters.

The scores from the second classifier (Table 2.4) are almost identical for all four prediction methods, with slightly higher Q_3, and a segment overlap score that is improves by between 1.5 and 5% over the first set of sequence-to-structure classifiers. The correlation coefficients also indicate that the second classifier has a slightly better prediction accuracy on all three classes.

Classifier	Q_3	Sov	SSEA	C_C	C_H	C_E	wrong
one-versus-rest	75.41	71.06	74.82	0.55	0.70	0.60	2.67
maxprob	75.36	70.63	74.87	0.55	0.70	0.60	2.78
one-versus-one	75.36	70.64	74.85	0.55	0.70	0.60	2.57
DAG	75.36	70.64	74.87	0.55	0.70	0.60	3.02
pair-wise couple	75.44	70.74	75.03	0.55	0.70	0.60	2.58
neural network	74.72	68.92	73.70	0.55	0.69	0.59	2.76

Table 2.4: Results from the structure-to-structure classifiers evaluated on a dataset of 75 test proteins. The SVM classifiers were trained on the pair-wise coupled probabilities. The neural network's scores are all significantly lower than the five SVM classifiers according to a paired t-test at the 95% level. None of the differences between the SVM classifiers are statistically significant according to the same criterion.

2.3.3 Estimating Classifier Accuracy using Cross-Validation

The final prediction method includes three 'one-versus-one' binary SVMs with the quadratic kernel in the sequence-to-structure layer, allowing training to be carried out on the full set of 1460 proteins because of the lower memory requirements. This approach is also roughly an order of magnitude faster to train than SVMs with the 'one-versus-rest' class assignments and the frequencies of each binary structure class are more evenly balanced. The outputs were mapped to probabilities and then fed into a second filtering layer of binary structure-to-structure SVMs with the Gaussian kernel. The error matrices of the cross-validated predictions are shown in Tables 2.5-2.7. The results are shown for the 1095 without unresolved chain breaks proteins out of the full set of 1460 . The average Q_3-score was found to be $77.07 \pm 0.26\%$ with a segment overlap score of $73.32 \pm 0.39\%$ and an SSEA score of $75.91 \pm 0.27\%$.

The SVM's $Q_3^{\text{obs}}(x)$ and $Q_3^{\text{pred}}(x)$ scores are quite similar to several other meth-

	H	E	C
obs(helix)	80.40	3.31	16.29
obs(sheet)	4.76	68.75	26.50
obs(coil)	10.63	10.15	79.22

Table 2.5: Classifier's assignment of the observed structural classes with diagonal entries representing the $Q_3^{\text{obs}}(x)$ scores for each structure type.

	H	E	C
pred(helix)	83.93	4.97	11.10
pred(sheet)	4.03	83.62	12.34
pred(coil)	13.35	21.71	64.93

Table 2.6: True class assignments of the predictions with diagonal entries indicating the $Q_3^{\text{pred}}(x)$ scores.

Q_3	Sov	SSEA	C_H	C_E	C_C
$77.07 \pm 0.26\%$	$73.32 \pm 0.39\%$	$75.91 \pm 0.27\%$	0.725	0.634	0.585

Table 2.7: Results from three-fold cross-validation of SVM on a data set of 1065 proteins. The confidence intervals in the Table above are given by $\sigma/\sqrt{n}$.

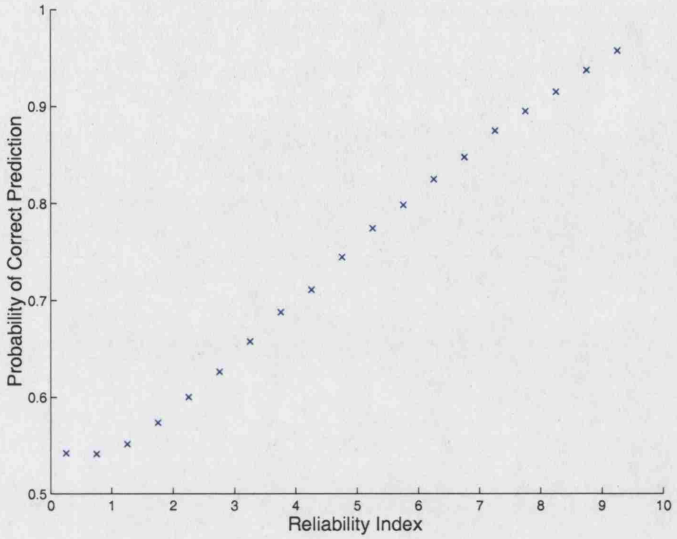

Figure 2.9: Reliability index for cross-validation set of 1065 proteins against posterior probability for bins of width 0.5.

ods with comparable accuracies, listed on the EVA server (Rost and Eyrich, 2001). The only slight differences are that the SVM is more conservative in predicting sheet residues, with a larger fraction misclassified as coil, but helices are predicted with slightly higher accuracy. The outputs from the first layer of binary classifiers give better estimates of the probability of a correct prediction than the outputs of the second layer. These estimates can be used to calculate a reliability index, as shown in Figure 2.9, which can be used to indicate the regions of a protein where the classification system has high confidence. The reliability score is simply given by the highest posterior probability minus the second-highest posterior probability multiplied by a factor of ten.

It should be pointed out that PSI-BLAST is able to detect homologous proteins at sequence identity thresholds lower than 25%, and that a match between a sequence and an existing structure at this level of similarity implies that the secondary structure of the sequence can be obtained accurately using homology with

the solved structure. This represents a situation where it would be more appropriate to use techniques other than secondary structure prediction, and consequently the accuracy of the prediction method, shown in Tables 2.2-2.7, may have been over-estimated. However, this effect is mitigated by the very large training sets which reduce the potential for simple 'memorization' of the training set. The following section overcomes this deficiency by comparing the SVM secondary structure prediction method with other state-of-the-art classifiers using a more stringent sequence similarity threshold.

2.3.4 Comparison of SVM Predictor of Secondary Structure with Other Modern Methods

Benchmarking of secondary structure methods is complicated by the great variation in accuracy between test proteins and the continual improvement in profiles resulting from the expansion of the sequence databases. As a result, an objective comparison of two methods can only be made on the same test set and with both methods having access to the same sequence database. This was achieved by benchmarking the SVM against PSIPRED on a set of 121 proteins released between January and June 2002. This set was filtered to remove any sequences that have homology with PSIPRED's training set (the SVM is trained on a subset) and each other.

The homology cut-off was set at a stringent PSI-BLAST expectation score of 0.1. The inputs were constructed using PSI-BLAST profiles from the same databases using identical search parameters with results shown in Table 2.8. The Table also includes results for the PROFsec prediction method (Rost and Eyrich, 2001) and a consensus of all three methods using a simple voting scheme.

The results of the SVM for this benchmark set of proteins are significantly lower than those estimated by cross-validation. However, both PSIPRED and PROFsec

	Q_3	Sov	SSEA	Length	Over	Under	Wrong
SVM	74.92	70.48	74.49	16.63	1.74	4.43	2.22
PSIPRED	74.97	71.81	74.32	16.37	1.85	4.30	2.50
PROFsec	74.90	71.15	73.94	16.27	1.63	4.88	2.28
consensus	76.17	72.37	75.41	15.75	1.74	4.08	2.23

Table 2.8: Accuracy scores for the SVM classifier compared to PSIPRED, PROFsec and a consensus of these methods on a test set of 121 proteins. The errors are divided into length errors, which occur at the ends of correctly predicted helix or sheet elements; over-predictions, where a coil segment is predicted as helix or strand; under-predictions, where helix or strand segments are predicted as coil; and wrong predictions, where strand is misclassified as helix and *vice versa*.

also achieve accuracies that are far lower than their overall averages (see Table 2.7) and the difference appears to be due to the difficulty of this particular test set. None of the differences in the scores between the SVM, PSIPRED and PROFsec are statistically significant at the 95% level according to a paired sample t-test. This suggests that the SVM has similar prediction accuracy to PSIPRED and the other top methods listed on the EVA server such as PROFsec, since common test sets of approximately 100 proteins have been sufficient to demonstrate a significant improvement over older methods such as PHD (Rost and Sander, 1993). Figure 2.10 also shows that the Q_3 scores for PSIPRED and the SVM on the 121 protein test set follow similar distributions.

The SVM therefore appears to have achieved similar prediction accuracy to PSIPRED despite being trained on 1460 rather than 5000 structures. The consensus achieves a significantly higher Q_3 and SSEA score than each individual method according to a paired t-test. This indicates that the SVM makes slightly different errors to PSIPRED and PROFsec, and that a consensus of the highest-accuracy

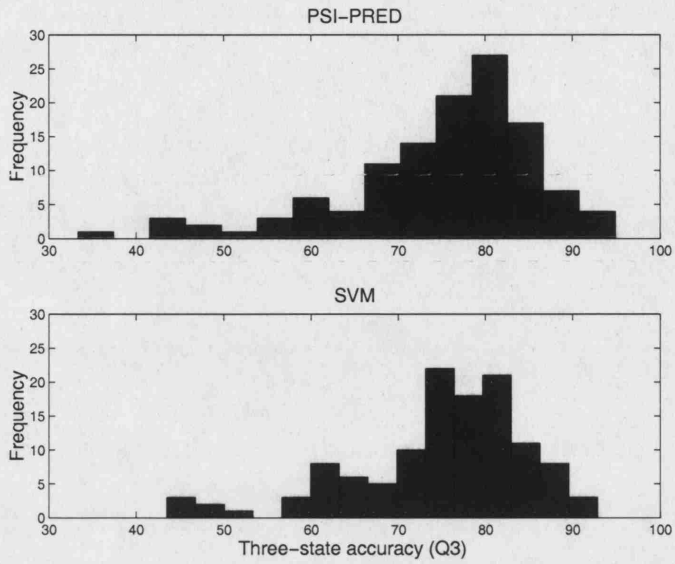

Figure 2.10: Histogram of Q_3 scores for PSIPRED and SVM for a set of 121 test proteins.

methods continues to improve on individual predictions (Cuff and Barton, 1999), even in the current era of structure prediction based on PSI-BLAST profiles. Most of the improvement occurs in predicting the ends of structural elements although there is also a reduction in the number of under-predictions and an improved SSEA score. The prediction accuracy of the support vector machine, estimated using cross-validation, is shown in Table 2.9 along with the accuracy scores for selected other methods from the EVA server.

The score for the SVM is significantly lower than that for PSIPRED on these large sets of proteins. However, the comparison is not made on common test sets and this difference could be explained by a systematic error such as PSI-BLAST failing to give an output for difficult targets submitted to the server or slightly different criteria for defining 'significant homology' between EVA and the present study.

Classifier	N	Q_3	Sov
SVM	1095	77.1	73.3
PSIPRED	1461	77.9	74.5
PROFsec	1554	76.7	73.0
SSpro2	1348	76.8	71.9
PHD	1599	70.9	66.9

Table 2.9: Accuracy scores for several prediction methods. N is the number of test proteins. The SVM has a prediction accuracy within a single standard deviation of the SSpro2 and PROFsec methods.

2.4 Discussion

The discussion is divided into two halves with the first half evaluating the success of the SVM algorithm for the secondary structure prediction problem. The second half discusses the future of secondary structure prediction in the wider context of research in structural bioinformatics.

2.4.1 SVMs in Secondary Structure Prediction

The support vector machine has demonstrated a similar accuracy to PSIPRED on the benchmark set of proteins, with a smaller training set and without use of ensemble averaging from several networks. The SVM also attains greater accuracy than a neural network when both classifiers are trained on small data sets but this advantage appears to be lost as the number of training examples is increased. The failure of a more advanced learning algorithm to improve secondary structure prediction is consistent with most of the recent developments in the field (Rost and Sander, 1993; Jones, 1999), which have shown that the generation of features that provide

a better representation of the conservation of residues in related structures is more crucial to the success of the prediction method.

The large training sets that are available to modern secondary structure prediction methods mean that the regularisation associated with maximising the margin does not greatly improve the generalisation error rate of the learning algorithm. Consideration of the error bound Equation 1 indicates that, for classifiers with a VC dimension[6] $d = O(10^2)$ that makes k mistakes on the $l = O(10^5)$ training examples, the error is dominated by the empirical error term (k/l), since, in this case

$$\text{err}(h) \leq \epsilon(l, H, \delta) = \frac{2k}{l} + \frac{4}{l}\left(d \log \frac{2el}{d} + \log \frac{4}{\delta}\right) \tag{2.9}$$

$$\epsilon(l, H, \delta) = \frac{2k}{l} + 0.03 \tag{2.10}$$

implying that overfitting is negated by the large number of examples.

A severe limitation on the utility of SVMs for this problem is caused by the linear growth in the number of support vectors with the size of the training set coupled with the high empirical error of secondary structure prediction, which means that the final method has a very large number of support vectors. Not only does the high fraction of examples which are support vectors compromise the generalization properties of the SVM, as noted earlier, but these support vectors occupy a total of 1.15Gb of memory and a single prediction requires 1.6×10^8 multiply-adds from the sequence-to-structure classifier alone. This computational complexity means that the SVM is not a practical alternative to neural networks for genome annotation or web servers. Although, there is significant redundancy in the set of support vectors and some simplification of the decision surface could yield a classifier with acceptable classification rates (Downs et al., 2001).

[6] linear classifiers have accuracies within a few percentage points of the polynomial classifiers.

This study has shown, however, that the uncalibrated SVM outputs can be used to provide estimates of posterior probability, and that a combination of these probability measures (Platt, 2000) and the pair-wise coupling technique (Hastie and Tibshirani, 1998) can be used to indicate confidence successfully. This approach may also be applicable to other areas of biological pattern recognition, although the sigmoid is not a good fit to the outputs of the second set of SVMs, indicating that the approximation does not apply in general.

It has also been shown that consensus predictions can be used to improve upon individual methods and an extension incorporating the most accurate modern prediction methods may achieve further gains in accuracy. However, most of the improvement occurs in correctly predicting the ends of structural elements and it is questionable whether this increased accuracy is valuable for use by either human experts or higher-level structure prediction methods.

Simplification of the SVM Decision Function

A major deficiency of non-linear soft margin SVM classifiers[7] is that the number of support vectors grows linearly with the size of the training set. This problem is acute for secondary structure prediction, which, in addition to large training sets, has a high empirical error and therefore a large number of bounded support vectors.

The rank of the kernel matrix K indicates the dimensionality of the set of support vectors in the high-dimensional feature space. Correlations between input vectors and the type of kernel function used often mean that the span of the support vectors is lower than the dimensionality of the feature space and lower than the cardinality of the set of support vectors. This means that the decision function, defined as a

[7]The primal weights of linear SVMs can be calculated explicitly and have the same dimensionality as the input vectors; see Section 3.5

weighted sum of the of support vectors, can be expressed as the sum of a subset.

It has already been shown that exact simplification of the decision surface is possible for simple problems (Downs et al., 2001). For larger problems, singular value decomposition (SVD) of the kernel matrix can be used to calculate matrix rank to a certain tolerance. Removing support vectors that are associated with singular values below the tolerance threshold can be used to calculate an approximate simplification of the decision surface. We have shown that this is possible for small and intermediate $O(10^3)$ numbers of support vectors (data not shown). However, the difficulty in computing the SVD for much larger kernel matrices precludes using this technique for many problems that occur commonly in bioinformatics. A more general solution would be to modify the objective function of the support vector machine to obtain more compact solutions, but this is beyond the scope of this thesis.

2.4.2 The Future of Secondary Structure Prediction

The most likely source of significant improvements in secondary structure prediction comes from algorithms that obtain a better representation of how structure is evolutionarily conserved or that make effective use of information from outside a relatively narrow input window. An approach aimed at utilising long range interactions might use the predictions from PSIPRED to estimate some initial constraints on the protein structure. These constraints could then be used to improve the predictions of the individual structural elements. For example, it may be possible to combine those regions of the protein with a high probability of being strand into sheet structures.

Since the publication of this work, it has been shown that refining sequence-based secondary structure predictions using the interactions between local residues in *de*

novo models of tertiary structure can improve the accuracy of both aspects of structure prediction (Meiler and Baker, 2003). This method, which uses the ROSETTA *ab initio* structure prediction algorithm, improves prediction accuracy dramatically on the relatively small number of proteins with Q_3 scores below 60%. These low scores usually arise from local regions of the sequence that are predicted to be helices but which form sheets as a result of interactions with other parts of the chain. Although the calculation of the models from ROSETTA is computationally intense, and the method is perhaps more appropriately categorized as an improvement of tertiary protein structure prediction, this does suggest that an increase of around 5% in accuracy can be achieved by incorporating longer range interactions in the prediction of secondary structure.

There are two other sources of error that apply to current secondary structure prediction methods which suggest that there is an upper limit on their potential accuracy. The first source of error, is mentioned in other parts of the text, and arises from the ambiguity of the class assignments at the ends of structural elements. These residues have some of the characteristics of regular structures but may, for example, not have the precise hydrogen bonding patterns of either sheets or helices. Results from the second CASP experiment suggest that ambiguous secondary structure assignments occur at a per residue frequency of around 5% (Lesk, 1997). The other source of error arises because of the variation in secondary structure between closely related sequences. The generation of profiles using PSI-BLAST or multiple alignments leads to an averaging of similar sequences, and as a result homologous but non-identical proteins have very similar profiles. The per residue variation of around 12% in the secondary structures between these proteins (Rost, 2001b) acts as another source of error for structure predictors that make use of sequence profiles.

This evidence, along with the absence of any significant improvement in the methods for predicting secondary structure in the previous five years, suggests that

three-state accuracies are approaching a theoretical upper limit of between 80 and 85%. Recent efforts have been made to predict the eight structural states, assigned by DSSP, using recurrent neural networks (Pollastri et al., 2002) and this may be a direction for future developments in the field. Some preliminary investigations that we have carried out with the SVM indicate that α- and 3_{10}-helices can be distinguished using evolutionary profiles, although this is complicated by the very unequal frequencies of the sub-divided structure classes. It is, however, questionable whether the eight-state predictions offer a great deal of practical benefit over the standard three-state predictions. A more useful improvement may arise from distinguishing between parallel and anti-parallel strands, which could aid assembly of super-secondary fragments into a complete structure.

In summary, this chapter has explored some of the practical problems that may be encountered when applying support vector machines to a well established problem in bioinformatics. The text describes solutions to obtaining posterior probabilities and combining binary classifiers. The time and space complexity of SVM classifiers is a well known deficiency of the algorithm that is yet to be resolved satisfactorily, although a partial solution has been discussed. The results on restricted data sets indicate that SVMs have the potential to outperform other standard machine learning techniques on problems where limited training data is available. The kernel function also provides a means for mapping discrete or non-Euclidean data representations such as amino acid sequences to a Hilbert space where linear discriminant functions can be found that separate the data. This advantage of kernel methods is discussed in later sections of this thesis.

Since the comparison of PSIPRED and the SVM algorithm was carried out, prediction accuracies have increased to around 79%, which is likely to be a consequence of improved sequence profiles being recovered from the ever-expanding sequence databases, perhaps with some small adjustments of existing methods. However,

the plateauing of prediction accuracies has led to a declining interest in secondary structure prediction, which is reflected by the omission of the secondary structure category from the sixth CASP experiment. This also reflects a slight decline in the importance of predicting globular protein structure that has arisen, in part, from the expansion of structural genomics projects.

The large structural genomics consortia have automated experimental protein structure determination, in a similar manner to the sequencing projects, by linking experimental groups to Laboratory Information Management Systems (LIMS) for selecting targets, processing data and sharing results between several sites (Burley, 2000). Improved technology has also significantly reduced the time taken to express, purify and crystallize proteins, and it seems likely that the vast majority of folds that can be solved readily using X-ray crystallography or nuclear magnetic resonance will be uncovered in the coming years. This has caused the focus of research in structural bioinformatics to shift to other areas such as the design of protein structures with novel topologies (Kuhlman et al., 2003) and functions, and to the aspects of protein structure that are more difficult to investigate experimentally such as the structures of large complexes (Aloy et al., 2004) and protein-protein interactions (Gray et al., 2003).

The following chapter describes a method for recognizing a class of protein structure that has, until recently, been largely ignored by the structure prediction community. This temporally disordered class of protein structure is difficult to investigate using X-ray crystallography and NMR spectroscopy, and appears to be important in key cellular processes such as transcription and cell signalling.

Chapter 3

Prediction of Intrinsic Disorder and Estimates of Disorder Frequencies in Complete Genomes

One of the central tenets of structural biology is that the function of a protein is determined by its three-dimensional structure. As a result, predicting protein structure has often been at the forefront of efforts to infer function (Laskowski et al., 2003; Pazos and Sternberg, 2004). However, it appears that a large proportion of protein sequences do not form completely globular structures. The *natively disordered* regions within these proteins may adopt an ensemble of structural states with transitions between the states leading to dynamic flexibility of the protein structure, or have non-globular structures that are extended in the solvent (Wright and Dyson, 1999).

It is well known that some degree of flexibility is present in many protein structures and that this flexibility is often essential for proper function. Most research into flexible structures has concentrated on either the small local movements caused by the 'induced fit' between the side chains of a protein and its ligand, or to the global 'hinge' or 'shear' movements of entire secondary structure elements or domains (Gerstein and Echols, 2004). However, it has begun to be accepted fairly recently that proteins in their native, functioning states can contain regions where the backbone atoms lack any stable conformation in solution, and that this dynamic flexibility is not some artefact of the experimental conditions such as the absence of an obligatory binding partner (Dunker and Obradovic, 2001).

The prediction of disordered regions could therefore provide a first step in identifying functionally important disordered regions such as those involved in molecular recognition and post-translational modifications (Iakoucheva et al., 2004). These disordered active sites may represent novel drug targets for the treatment of diseases such as cancer (Iakoucheva et al., 2002). Disorder has also been implicated in prion diseases (Donne et al., 1997), and it is now known that some disordered regions are involved in the formation of the β-sheets between chains which initiate aggregation and eventually amyloidosis (DuBay et al., 2004). Disorder prediction is also

proving to be a valuable tool for structural genomics projects where the removal of unstructured regions is often vital for the successful crystallization of proteins prior to X-ray structure diffraction studies.

The premise that structure is determined by primary sequence might also be applied to lack of structure or disorder. There are also clear patterns that characterize disordered regions such as low sequence complexity, amino acid compositional bias (e.g. toward charged residues) and high flexibility, and it has been shown in a series of papers (Romero et al., 1997; Li et al., 1999; Romero et al., 2001; Dunker and Obradovic, 2001) that disordered regions can be predicted successfully from amino acid sequence.

This chapter reviews the latest disorder prediction methods, and describes the development of several new classifiers. The first classifier, DISOPRED, is based on a feed-forward neural network, and was assessed at the fifth Critical Assessment of techniques for Structure Prediction (CASP) experiment, which included an evaluation of disorder prediction methods (Melamud and Moult, 2003). The second classifier, DISOPRED2, was trained using a support vector machine (SVM) learning algorithm and is benchmarked on the same targets. Several outstanding questions arising from CASP5, such as the extent to which information from homologous sequence improves prediction of disorder, were also addressed by developing several other classifiers, which were trained and evaluated on identical data sets.

This chapter also describes the use of DISOPRED2 to investigate the frequency of disorder in several archaea, eubacteria and eukaryote genomes. Previous genome-wide analyses of disordered regions have been based on classifiers with high false positive rates (the classifier developed by Vucetic et al. (2003) had a false positive rate of $\sim$ 16% for disordered segments longer than 40 residues). Although the results presented here cannot be interpreted as a lower bound on the proportion of proteins

that contain disorder, they are intended to be very conservative with false positive rates expected to be lower than 0.5% on long disordered segments.

The final section describes the design and implementation of the DISOPRED2 server, which provides a web interface to the DISOPRED2 prediction method. The server has received an average of around 50 submissions per week since it was released in April 2004. An executable version of the DISOPRED2 software is also being used by NMR and crystallography groups. The following section describes some of the experimental techniques that have been used to define disorder in protein structures.

3.1 Experimental Techniques for Investigating Native Disorder

Several experimental techniques can be used to identify native disorder in protein structures. In this chapter, the experimental definition of disorder comes from highly-resolved X-ray crystal structures, and this definition is discussed in greater depth in subsequent sections. However, there are various other techniques from spectroscopy and molecular biology that have been used to probe disorder in protein structures. The more commonly-used techniques are circular dichroism (CD), nuclear magnetic resonance (NMR) and proteolytic degradation (PD), and these are discussed in greater depth in this section. The DisProt database also includes examples of natively disordered structures that have been characterized using mass spectroscopy, electron microscopy and infra-red spectroscopy, and indirect molecular biological techniques such as immunochemistry and gel filtration (Vucetic et al., 2005). These are typically limited to fewer than three proteins, and are not discussed in greater detail.

Circular dichroism spectroscopy passes plane polarized light in the far UV spec-

trum through diluted protein solutions. The plane polarized light can be viewed as a superposition of opposite circularly polarized light of equal amplitude and phase. The regular structural elements have different absorbances for left- and right-handed circularly polarized light, resulting in *ellipticity* of the resultant wave. Helices and strands cause ellipticity in the incident wave at different frequencies, so the absorption spectrum of the purified and diluted protein can be used to measure the overall secondary structure content of the protein. Spectra indicating a low proportion of helix and sheet elements are often interpreted as evidence of disorder. However, there are several weaknesses to this approach. Firstly, CD spectroscopy is inaccurate, and is slightly poorer than secondary structure prediction for determining the overall structural class of the protein (Rost, 1996). Secondly, CD spectroscopy measures the global properties of the protein, and cannot be used to identify local regions of order/disorder. And thirdly, an absence of regular secondary structure elements does not necessarily indicate that the protein is disordered (Liu et al., 2003).

Nuclear Magnetic Resonance is another spectroscopic technique that can be used to determine protein structure and investigate protein dynamics. NMR uses several quantum mechanical effects arising from the magnetic moment or spin of certain nuclei. The spectra, which are obtained by passing radio frequency waves through a sample subjected to a strong magnetic field, can be used to infer nuclei-nuclei distances in the protein structure. These distances act as constraints for constructing the structural model of the protein. It is often the case that the data obtained from NMR experiments is incomplete, which leads to several structural models fitting the distance and torsion constraints equally well. However, insufficient constraints can also arise from the protein being disordered in solution (Branden and Tooze, 1999). It is possible to remove this ambiguity by the use of spin relaxation methods, which can be used to resolve protein motions on pico- and nano-second time scales (Bracken, 2001).

Proteolytic degradation is one technique from molecular biology that can be used to detect disorder indirectly. This technique is based on the principle that disordered regions are cleaved more readily by proteases than globular portions of the protein (Vucetic et al., 2005). The cleavage sites can be identified by mixing proteases with the purified protein and performing gel electrophoresis on the resulting fragments. Experimental molecular biology and techniques such as PD and immunochemistry are useful for providing further verification of the disorder/order regions that have been established by other means.

The following section describes previous work on learning algorithms for predicting disorder and the use of the resulting classifiers for estimating disorder frequencies in complete genomes.

3.2 Review of Computational Approaches to the Prediction of Native Disorder from Amino Acid Sequences

The Dunker-Obradovic groups were the first to show that machine learning algorithms could be used successfully for local disorder prediction based on amino acid sequences. Their initial method used a sequential forward search algorithm to select the net charge, hydrophobicity and the frequencies of the amino acids R, D, E, K, F, W, Y as features for predicting disorder. These features were calculated for windows of 21 residues in length, and used to train feed-forward neural networks (Romero et al., 1997). The experimental definition for the disordered residues came from long, internal regions of disorder discovered using X-ray crystallography and NMR. This method was later augmented with a similar predictor for the N- and C-termini (Li et al., 1999) to form the method VLXT.

VLXT was followed by a method using ensembles of linear least-squares classifiers, which were trained with a larger and more carefully prepared data set (VL2). The classifiers were trained using a competitive learning strategy to partition the training set into "flavors" of disorder that were characterized by distinct amino acid compositions and functions (Vucetic et al., 2003). The latest method (VL3) is an ensemble of three multi-layer perceptrons, trained to partition the training set. The features used by both VL2 and VL3 are amino acid composition, average flexibility and average sequence complexity of a window of 41 residues. The authors entered these and several other prediction methods in the fifth CASP experiments but the results suggest that any improvements in performance over their earliest classifier (VLXT) are, at best, moderate (see Section 3.3.4).

The Dunker-Obradovic collaboration has also published several estimates of the frequency of disorder in complete genomes (Dunker et al., 2000; Dunker and Obradovic, 2001; Vucetic et al., 2003). The most comprehensive of these studies used the VLXT predictor to estimate disorder frequencies in a set of 34 complete genomes, including 7 archaea, 22 eubacteria and 5 eukaryotes (Dunker et al., 2000). The false positive rate for the VLXT predictor was estimated by applying the method to a non-redundant (25% sequence identity) set of ordered PDB structures. This gave a per chain false positive rate for long (> 30 residue) disordered segments of 17%. This very high false positive rate is undesirable, since the 'true' rate could be significantly higher because of the biased nature of the sampling (e.g. successful crystallization) used to create the PDB. Although, this limitation was recognized by Dunker et al. (2000), no attempt was made to minimise this potential source of error.

When applied to complete genomes, VLXT predicted $37 \pm 7\%$ of archaea, $30 \pm 2\%$ of eubacteria and $54 \pm 3\%$ of eukaryote protein chains to contain a region of disorder with length greater than 30 residues. However, there was significant variation

between species within each kingdom, with disorder estimates ranging between 9 and 53% in the archaea, 14 and 52% in the eubacteria and 48 and 63% in the eukaryota. There are several other anomalies in the results, such as the large differences in disorder frequencies that are observed between related organisms. The most striking example is the difference in the frequency of long, predicted, disordered segments between *Campylobacter jejuni* (14%) and *Pseudomonas aeruginosa* (42%) since these species are both members of the proteobacteria. Conversely, unrelated and morphologically divergent organisms such as the nematode *Caenorhabditis elegans* and the archaea *Halobacterium* Sp. have similar frequencies (49% compared with 53%). This article was acompanied by a commentary in *Nature Biotechnology*, which claimed that protein structures existed in a dynamic equilibrium between a 'trinity' of states; ordered, collapsed and extended (Dunker and Obradovic, 2001). The commentary also suggested that the disordered sequences were associated with signalling cascades and the ribosome, and that disorder might have been a prerequisite for multicellularity.

The VL2 method for partitioning disorder into distinct 'flavors' was also used to provide estimates of disorder frequencies in complete genomes (Vucetic et al., 2003). However, the classifier also had a high false positive rate (16% for segments of length 40 or greater), and provided similar estimates to the previous study using VLXT. This work also provided tentative evidence for each 'flavor' of disorder being associated with specific structures and functions.

Another research group working on native disorder (Uversky et al., 2000) developed a simple method based on the net charge and mean hydrophobicity of the protein sequence to obtain a global prediction of whether a protein adopts a globular structure. The non-globular structures were operationally defined as random-coil conformations determined from NMR experiments or a lack of significant ordered structural elements as determined by CD spectroscopy. Similar calculations are used

to predict local regions of disorder using a sliding window by the FoldIndex server (Prilusky et al., 2003).

Other work has been carried out using neural networks for the prediction of extended regions with no regular secondary structure (NORS), which are defined as segments of 70 consecutive residues with less than 12% helical or strand content (Liu et al., 2003). NORS segments are not necessarily disordered but have similar compositional biases and sequence complexity to disordered regions. The finding that around 20% of eukaryote proteins contained a NORS region, accounting for 15% of residues, but that there are far fewer NORS regions in prokaryotes, suggests that many "loopy" regions are also predicted as disordered. This is supported by the lower number of hydrogen bonds formed by residues within NORS regions (0.66) compared with non-NORS regions (1.21). The observation by Liu et al. (2003) that NORS regions are as conserved as flanking regions, and that many NORS segments are involved in molecular recognition is also consistent with the properties of disorder described later in this chapter.

Another recent contribution to the field used back-propagation networks to predict disorder according to two separate definitions (Linding et al., 2003). One definition used the missing residues from X-ray crystal structures and the other normalised B-factors, which represent the degree of thermal motion of specific atoms in the structural model. The rationale for using missing co-ordinates is that, if a series of residues are disordered in solution, they will not adopt identical conformations in the protein structures that form the crystal. Consequently, this region of the protein will not scatter X-rays coherently and will appear as a diffuse area of electron density. The crystallographer will then be unable to assign a conformation to the back-bone or the orientations of the side-chains, and typically will not include these co-ordinates in the structural model. However, there is no single systematic procedure for determining structural models, and it is possible that regions of disor-

der in an electron density map could be interpreted by different crystallographers as regions of the model with either missing co-ordinates or several conformations with low occupancy (Branden and Tooze, 1999).

It is speculated that residues with high B-factors or "hot loops" have some of the properties of disorder, such as increased flexibility. However, high B-factors are often associated with highly motile side chains that are exposed to the solvent rather than the backbone atoms. For example, both lysine and proline are enriched in positions with missing co-ordinates (see Figure 3.4) but whereas lysine is also common in "hot loop" segments, proline is under-represented in residues with high B-factors (Linding et al., 2003). This is a presumably a consequence of lysine's long, charged side-chain having much greater conformational freedom than proline's side-chain with its rigid ring structure. Despite this, predictors trained on the two definitions of disorder do have some correlation (C=0.46), which may arise from mobile side-chains being a necessary but not sufficient property of disordered structures, i.e. natively disordered regions necessarily have unconstrained side-chains as a consequence of the flexible backbone and high solvent exposure, but high B-factors can also occur in static loop regions.

3.3 System and Methods

The training set for DISOPRED2 was the same as that used to train the original version of DISOPRED (Jones and Ward, 2003) and was composed of non-redundant chains with X-ray structures in the Protein Data Bank (Berman et al., 2000) and less than 25% pair-wise sequence identity. This ensured that a large, non-redundant training set was available to the classifier and that cross-validation experiments would provide estimates of the prediction accuracies that would be expected for proteins that do not have high sequence similarity with existing structures.

Only structures with resolutions better than 2.0 Å were used to ensure that missing regions were not caused by poor overall model quality. The disordered residues were identified by aligning the sequence of the protein chain in the SEQRES records with the sequence as specified by the ATOM records (alpha-carbon coordinates). Residues which were found in the SEQRES records but not in the ATOM records (gaps in the alignment) were defined as disordered. The final training set comprised 715 protein chains, in which a total of 176550 residues were defined as ordered and 4590 residues as disordered (disordered residues account for 2.53% of the training set). For each protein in the training set, a sequence profile was generated using three iterations of a PSI-BLAST search against a large, non-redundant sequence database (Holm and Sander, 1998). The threshold for including hits in the calculation of the position-specific scoring matrix for subsequent rounds of the search was set at an expectation value of 10^{-3}. The sequence database was filtered to mask transmembrane regions but not low complexity or compositionally biased regions, as these properties are associated with disorder (Romero et al., 2001).

The window length was set to fifteen residues, as preliminary investigations with other window sizes indicated that accuracies were optimal in the range between around 10 and 18 residues. The length distribution of the disordered segments in the training set for DISOPRED2 provides a justification for this choice of window size, since the majority of disordered regions (90% of disordered segments accounting for 61% of the disordered residues) have lengths shorter than 15 residues (see Figure 3.1). This is also an identical length to that chosen for secondary structure prediction (Chapter 2), and would also be expected to allow the classifier to distinguish disorder from the regular structural elements.

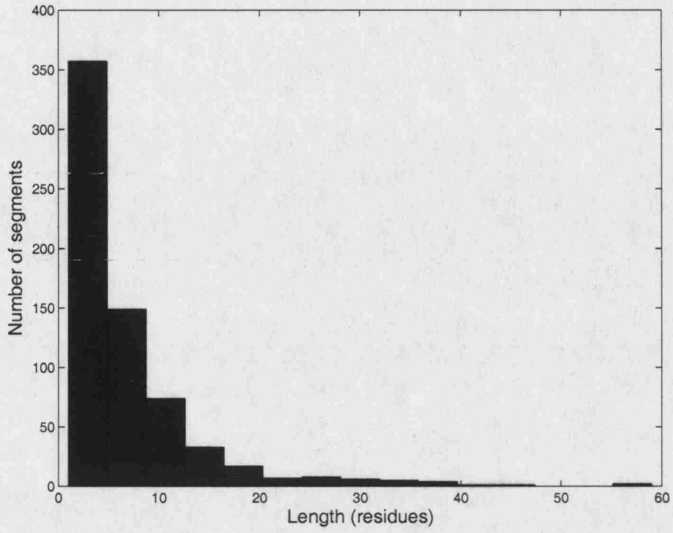

Figure 3.1: Length distribution (in residues) of segments missing from the electron density map of highly resolved crystal structures. Two segments that have lengths greater than 60 residues are omitted from the distribution (these segments are 78 and 82 residues in length).

3.3.1 Learning Algorithms for Recognizing Native Disorder in DISOPRED and DISOPRED2

The first version of DISOPRED was based on a feed-forward neural network with a very similar topology to PSIPRED. This network had 315 input units from the twenty profile scores for each position in the window of fifteen residues, with an extra input unit used to indicate whether the window extended beyond the N- or C-termini. The profile scores were normalized to the range $[0, 1]$, similarly to PSIPRED (Jones, 1999), in order to decrease the time taken to train the network. The network included a single hidden layer containing 55 neurons and an output layer containing two outputs, and was trained using online back-propagation with a momentum term. Overfitting was controlled by stopping training when performance began to decline on a validation set, which contained 10% of the training data. Predictions of disorder that coincided with confident PSIPRED predictions of helix

or sheet were then manually assigned marginal predictions of order for the purposes of the CASP5 experiment (Jones and Ward, 2003).

Chapter 2 indicated that support vector machines outperformed feed-forward neural networks when the training set was restricted. In this case, although the training set is relatively large, the number of examples of disorder is small compared with the dimensionality of the input space (4590 compared with 315). This suggested that controlling the capacity of the resulting classifier by maximising the margin may improve generalization over a neural network trained with weaker capacity control. The following section compares the accuracies of disorder predictors trained using the SVM and a neural network learning algorithm, and addresses several potential deficiencies of the original method, such as the reliance on manual adjustment of classifier outputs, the treatment of the terminal positions, and the absence of a second cascaded classifier.

The *SVMlight* support vector machine package (Joachims, 1999) was used to train DISOPRED2. Unbalanced class frequencies can result in classifiers that output the majority class exclusively since this optimises overall accuracy. This behaviour is prevented in a formulation of the SVM that places asymmetric costs on points that violate the geometric margin (Morik et al., 1999). This allows a greater cost to be placed on margin breaches by points from the minority (disordered) class than examples from the majority (ordered) class. Correctly setting the asymmetric cost parameter results in informative classifier outputs.

The asymmetric cost is incorporated into the SVM learning algorithm by replacing the single slack variable term in the 1-norm soft-margin objective function (Equation 1.20), by separate terms for the positive and negative examples

$$\begin{aligned} \text{minimise}_{\mathbf{w},b} \quad & \langle \mathbf{w} \cdot \mathbf{w} \rangle + C_+ \sum_{i:y_i=1} \xi_i + C_- \sum_{j:y_j=-1} \xi_j \\ \textit{subject to} \quad & y_k(\langle \mathbf{w} \cdot \mathbf{x}_k \rangle + b) \geq 1 - \xi_k \\ & k = 1, \ldots, l, \end{aligned} \tag{3.1}$$

where C_+ is the cost on margin breaches by examples in the positive class and C_- the cost for negative examples. In the dual form of the objective function, this results in separate 'box' constraints on the Lagrange multipliers associated with examples in the positive and negative classes.

Preliminary investigations with various standard kernel functions (linear, polynomial, Gaussian) indicated that the highest accuracy was achieved with the linear kernel. The learning parameters (regularization parameter and cost parameter), were found using an exhaustive search using 4-fold cross-validation on the training set. The relative performance of the classifiers was measured using Receiver-Operating Characteristic (ROC) curves and associated statistics from the pooled results, as described in the following section.

3.3.2 Benchmarking SVM Performance using Receiver Operating Characteristic Curves

The ROC curve displays a plot of how the fraction of recovered positive examples increases as the decision threshold is relaxed to allow more false positives, and is particularly appropriate for problems where there is an imbalance in the class frequencies, and high accuracies can be obtained trivially (Duda et al., 2000). The ROC curve is also useful for comparing the performance of two classifiers over the entire range of possible false positive rate thresholds.

The area under the ROC curve A_{roc} has a simple relationship to the Mann-Whitney statistic, U, and the Wilcoxon rank-sum statistic, T, (Rees, 1995).

$$A_{roc} = \frac{U}{mn} = \frac{1}{mn}\left(T - \frac{m(m+1)}{2}\right) \tag{3.2}$$

where m and n are the numbers of examples in the positive and negative classes, respectively. These are used in non-parametric tests of whether two samples are taken from the same distribution and can be used to determine which of the two rankings supplied by different classifiers provide the better separation of positive and negative examples. The definitions of the two statistics for binary pattern recognition problems is given below

Definition 1 *Let V be the set of m functional outputs $f(\mathbf{x})$ from the positive distribution $(\mathbf{x}_i, y_i) \in X \times 1$ and W the set of n outputs from the negative distribution $(\mathbf{x}_j, y_j) \in X \times -1$ such that when all the observations are ranked, V_i takes rank R_i and W_j takes rank S_j (no ranks are tied). The rank sum statistic is $T = \sum_{i=1}^{m} R_i$ and the statistic U is given by*

$$U = \textit{number of pairs } (V_i, W_j) \textit{ with } V_i > W_j \tag{3.3}$$

An A_{roc} score of 0.5 therefore represents random and 1 perfect classification. Since the relationship between the Wilcoxon statistic and the area under the ROC curve is linear, the two terms are used interchangeably in the rest of this section.

The standard error calculation for ROC curves on a particular test set has been developed from work on medical diagnosis (Hanley and McNeil, 1982), and assumes a normal distribution for A_{roc} scores. However, the estimate is conservative, and does not depend on the particular distribution of positive and negative examples

$$SE = \sqrt{\frac{A_{roc}(1 - A_{roc}) + (n-1)(Q1 - A_{roc}^2) + (m-1)(Q2 - A_{roc}^2)}{mn}} \quad (3.4)$$

where the quantities $Q1$ and $Q2$ are given by

$$Q1 = A_{roc}/(2 - A_{roc}) \quad (3.5)$$

$$Q2 = 2A_{roc}^2/(1 + A_{roc}) \quad (3.6)$$

The Wilcoxon test statistics can also be used to indicate whether the difference between two methods is statistically significant. The standard error of the difference in A_{roc} scores between two methods applied to the same test set (A_1 and A_2) is given by

$$SE(A_1 - A_2) = \sqrt{SE^2(A_1) + SE^2(A_2) - 2r.SE(A_1)SE(A_2)} \quad (3.7)$$

where r is the correlation coefficient between the functional outputs of the two methods (Hanley and McNeil, 1983).

3.3.3 Investigating the Utility of Evolutionary Information for Predicting Disorder

This step was carried out to investigate whether profiles containing evolutionary information improved the accuracy of disorder prediction over that obtained from single sequences. The secondary structure predictions were included to test whether explicit predictions of the regular helix and sheet regions improved recognition of disorder. Separate classifiers were trained for the N- and C-termini, as it has been

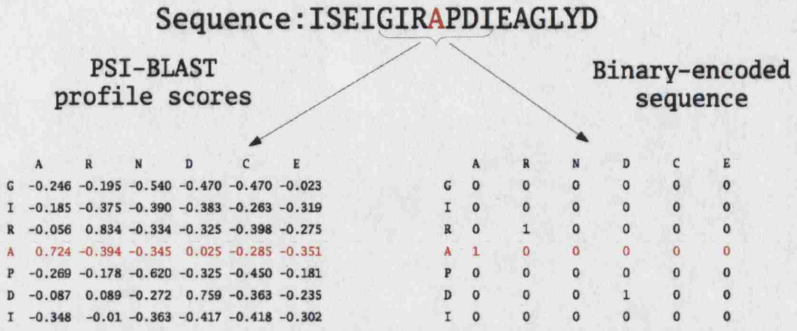

Figure 3.2: Procedure for encoding amino acid sequences, PSI-BLAST profiles and PSIPRED predictions for predicting native disorder. A window of seven residues is encoded using a binary representation of the amino acid sequence and the profile from three iterations of a PSI-BLAST search. These matrices are concatenated into a single vector for training the classifiers. The PSIPRED predictions are included as three extra inputs (coil, sheet and helix) for each residue in the window.

demonstrated previously that there are slightly different patterns in disordered sequences at the terminal positions (Li et al., 1999) but these were found to have little effect on the overall prediction accuracy.

The diagram in Figure 3.2 shows how the amino acid sequences, PSI-BLAST profiles and PSIPRED predictions were converted into feature vectors for supervised learning. The profile scores for each amino acid were scaled to have a mean of zero and a variance of one. The transformation reduces the potential for loss of numerical precision in solving the quadratic program (Golub and van Loan, 1996).

Figure 3.3 shows Receiver Operating Characteristic (ROC) curves for several SVMs trained using combinations of binary-encoded amino acid sequence, secondary structure predictions from PSIPRED (Jones, 1999) and profiles from iterated BLAST searches (Altschul et al., 1997) for symmetric windows of fifteen positions. Table 3.1 shows estimates of the area under each ROC curve for the four classifiers shown

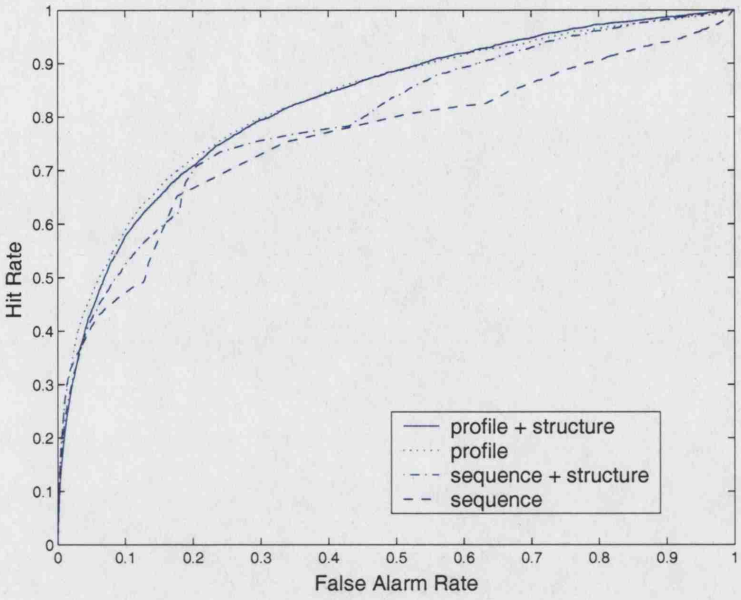

Figure 3.3: Receiver Operator Characteristic curves for linear SVM classifiers generated using four-fold cross-validation on the non-redundant set of proteins. The ROC curves were generated by varying the decision threshold of each SVM classifier.

in Figure 3.3, in addition to the results from a second smoothing classifier trained on the outputs of the profile SVM. This classifier was constructed by mapping the outputs from the profile classifier to posterior probabilities using a logistic sigmoid function (Platt, 2000). The posterior probabilities from a window of 15 residues was then used to used to train a linear SVM classifier to predict the disorder/order state of the central residue. In this case, all the differences apart from the two profile-based classifiers are significant at the 95% level according to the test described previously (Hanley and McNeil, 1983).

Classifiers trained on PSI-BLAST profiles outperform those trained on single sequences across the range of error rate thresholds indicating that evolutionary information improves the prediction of disorder. Secondary structure predictions appear to improve the accuracy of the sequence classifiers because they contain implicit

	MCC	Q_2	Wilcoxon	SE
profile + structure	0.257	93.74	82.70	0.34
profile	0.274	93.79	83.02	0.36
sequence + structure	0.246	93.69	79.84	0.38
sequence	0.237	93.63	76.16	0.46
cascaded classifier	0.350	94.05	86.75	0.30

Table 3.1: Table shows the Matthew's correlation coefficient and two-state accuracy (Q_2) for a false hit rate of 0.05 and the Wilcoxon statistic with its standard error (SE). The area under the ROC curves (Wilcoxon statistic) was calculated using trapezium rule numerical integration.

information from the position-specific scoring matrix but do not improve the profile classifiers. This contradicts the results from the first version of DISOPRED (Jones and Ward, 2003), where structure predictions were used to improve accuracy.

The difference may be attributable to the different learning algorithms used in the two cases. This study uses a linear SVM with a relatively low capacity (Vapnik, 1998) (i.e. it will avoid overfitting the data) and a capacity-controlling maximal margin learning algorithm. On the other hand, the two-layer neural network, used to train the original classifier, had a relatively large number of hidden units and may have been prone to overfitting. It is possible that the improved generalization of the SVM prevents prediction of disorder in regions that are likely to form helical or strand elements in the core of a globular protein. The linear profile classifier is referred to as DISOPREDsvm in subsequent sections of this chapter.

There are also several experimental examples where regions of the protein that contain regular structural elements are found to be partially disordered (Daughdrill et al., 1998). Most of these elements appear to be helical and can arise from a rigid

helix element connected to a globular protein by flexible loops (Figure 3.9) or a region where the disorder and helix states are in equilibrium. These regions often undergo a transition to the helical state upon binding to another protein or DNA (Parker et al., 1999).

Although evolutionary information improves the prediction of disorder very slightly, the gain in accuracy is not as great as in secondary structure, where predictions based on profiles achieve three-state accuracies that are at least 15% higher than those made on single sequences. An explanation for this may be that profiles greatly improve secondary structure prediction by implicitly encoding constraints on the protein structure outside a relatively narrow input window, whereas dynamically disordered structures, by definition, are not subject to such constraints prior to being stabilized by some other interaction.

Further improvements in accuracy can be achieved by inputting the first set of predictions into a second cascaded network, which extends the effective length of the input window from 15 to 29 residues and increases prediction accuracies of longer (> 15 residue) disordered segments. The cascaded classifier therefore has the familiar two-layer neural network topology with fifteen hidden units that have a sigmoid activation function but without full connectivity in the first layer of adaptive weights. DISOPRED2 is comprised of this type of cascaded classifier, trained on the full set of proteins.

3.3.4 Comparison of DISOPRED2 and DISOPREDsvm to other Disorder Recognition Algorithms on the Targets from CASP5

An objective comparison was carried out between DISOPRED2 and several other disorder prediction methods evaluated on targets from the fifth CASP experiment. The PSI-BLAST search database and the training set for DISOPRED2 and DISO-

PREDsvm were compiled before the start of CASP so the test can be considered fair.

The Dunker-Obradovic groups have claimed that predictions based on sequence composition produce higher accuracies than those obtained using amino acid sequences directly (Romero et al., 1997; Li et al., 1999; Romero et al., 2001). This hypothesis was tested by constructing a simple predictor based on the average propensity of each amino acid residue (A) for forming disorder (D).

$$P_A(D) = \log\left(\frac{p(A|D)}{p(A)}\right) = \log\left(\frac{n_{D,A}}{n_A n_D/n}\right) \tag{3.8}$$

where n_x is the number of residues in state x, giving the log-odds for each amino acid being associated with disorder. The propensities (shown in Figure 3.4) were averaged over a window of fifteen amino acids to provide an order/disorder score for the central residue. This classifier is similar to the original Chou and Fasman (1974) method for the prediction of secondary structure and is referred to as DISOcf in the remainder of the text.

Figure 3.4 indicates that the hydrophobic residues (particularly those with aromatic side-chains) are under-represented in disordered regions. The common occurrence of cysteine in enzyme active sites and its ability to form covalent disulphide bridges may explain its propensity toward forming ordered structures. The basic residues (R, H, K) and serine are disposed to forming disordered secondary structures. The high propensity of methionine for forming disorder is likely to be caused by its tendency to occur at the N-terminus. An identical trend in amino acid propensities has been observed in a previous study of disordered sequences by the Dunker group (Romero et al., 2001).

Figure 3.5 and Table 3.2 show results from DISOPRED2 and DISOcf, along with

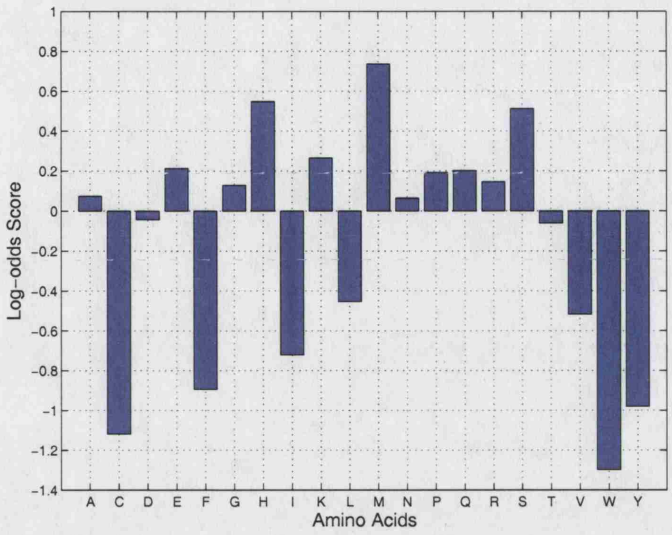

Figure 3.4: Amino acid propensities for forming disorder. The single letter amino acid code is used along the x-axis. Scores greater (less) than zero indicate residues that are associated with disordered (ordered) structures.

those from the linear profile SVM (DISOPREDsvm) and the neural network version of DISOPRED with the manually adjusted residues removed (DISOPREDnn). Results from the Obradovic (VL3) and Dunker (VLXT) groups submitted in the model 1 category and the VL2 method from the Dunker group, which achieved highest accuracy according to their own assessment (Obradovic et al., 2003) are also shown. The FoldIndex program was not entered in the fifth CASP experiment but was included in this comparison as a simple predictor, based on only mean net charge and hydrophobicity of the sequence (Prilusky et al., 2003). The window parameter for the FoldIndex program was set to 31 residues, as this value achieved highest accuracy on a validation set.

The DISOPRED predictors achieve higher accuracies than the other methods across the whole range of decision thresholds apart from a slight deficiency over VL3 at very low false positive rates ($< 1.27\%$). Applying a simple rule to the

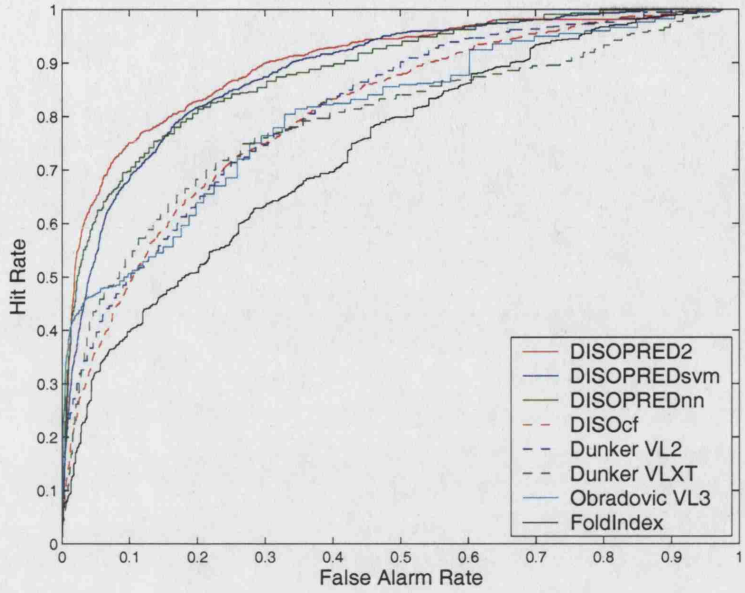

Figure 3.5: Receiver Operator Characteristic curves comparing the outputs of DISOPRED2 to six other methods evaluated on the targets from CASP5.

	MCC	Q_2	precision	recall	Wilcoxon	SE
DISOPRED2	0.511	93.1	46.4	64.6	90.02	0.64
DISOPREDsvm	0.437	92.4	42.1	54.2	88.58	0.62
DISOPREDnn	0.431	93.6	35.1	61.0	88.63	0.82
DISOcf	0.301	91.2	33.3	36.2	80.72	0.79
Dunker VL2	0.355	91.8	36.8	43.3	79.08	0.97
Dunker VLXT	0.313	91.4	33.9	38.2	81.31	0.78
Obradovic VL3	0.382	92.0	38.6	46.8	80.59	0.88
FoldIndex	0.262	91.0	30.1	32.0	73.88	0.92

Table 3.2: Table shows the Matthew's correlation coefficient (MCC), two-state accuracy (Q_2), precision and recall for a false alarm rate of 0.05 and the Wilcoxon statistic and its standard error for the targets from CASP5.

outputs of the DISOPRED predictors that removes predictions for short disordered segments yields even higher accuracies at the low thresholds. Since this property applies whichever learning algorithm is used to train the DISOPRED classifiers, it is likely that the improvement is due either to the definition of disorder used in training or the means of encoding the protein sequence.

Table 3.2 shows clearly that the means of encoding the protein sequence has the greatest effect on the accuracy of disorder prediction, as the three most accurate methods were trained using input features that were derived directly from the amino acid sequence. (In this case sequence profiles, although the previous section indicated that similar accuracy could be achieved with single sequences). The methods trained using various properties that are calculated from amino acid composition (DISOcf, VLXT, VL2 and VL3), had prediction accuracies that were very similar to each other and substantially lower than the DISOPRED classifiers. The amino acid frequencies selected for disorder prediction in the development of the VL2[1] method are also similar to the residues that have high propensities for forming either order or disorder, which suggests that the training sets for the two methods do not differ greatly. The FoldIndex program makes predictions on only two features, the hydrophobicity and net charge, and has the poorest performance of the eight disorder predictors.

The similar results for DISOPREDsvm and DISOPREDnn also indicate that the SVM does not greatly outperform a comparable neural network for this problem, although two cascaded SVMs (DISOPRED2) do have significantly higher prediction accuracies (at the 95% level according to the test described previously). Several features of this problem, such as the class imbalance and high empirical error, remove any learning-theoretic bounds that justify maximising the margin. The VC dimension of the hypothesis class used to obtain the decision function for disorder

[1]The amino acids: A, F, I, L, V, W, Y

prediction is also low[2], and is unlikely to lead to overfitting.

The differences in the Wilcoxon scores between the DISOPRED2 predictor and the other seven methods are all statistically significant at the 95% level. This suggests that the DISOPRED2 classifier is more accurate than other published disorder predictors. (The disEMBL method developed by Linding et al. (2003) did not compete at CASP5, but recovered slightly fewer (0.45 compared with 0.57 in the present study) disordered regions at a false positive rate of 0.05 in their own benchmarking experiments, although more stringent homology cut-offs were used for cross-validating their results.) The success of the DISOPRED2 classifier motivated us to investigate the frequency of disorder in completed genomes to gain a greater knowledge of the potential limitations of structural genomics projects, and of the function and evolution of disorder, as described in the following section.

3.4 Predicting the Frequency of Disorder in Complete Genomes

The disorder frequency in sequenced genomes was estimated by first applying DISOPRED2 to a non-redundant set of ordered protein structures to obtain an estimate of the false positive rate of the classifier. The decision threshold was set conservatively (a per residue rate of 2%) to ensure that the estimates could be considered fairly confidently to be a lower bound on the disorder frequency in Nature. At this threshold, around 60% of long, predicted regions of disorder are found to coincide with long regions of the sequence that have missing co-ordinates in crystal structures.

Although DISOPRED2 was developed with the aim of optimising per residue accuracy, it is important to distinguish between long contiguous regions of disorder

[2]The set of affine functions in the input space with dimensionality 315.

and short disordered segments, which are less likely to be functionally relevant. For this reason, the fraction of protein chains that contain long disordered regions (> 30 or > 50 residues) was also calculated.

3.4.1 Estimating False Positive Rates using Ordered Crystal Structures

The false positive rate for DISOPRED2 was established by classifying a set of 7169 ordered proteins (all residues for the protein set have atomic co-ordinates recorded in the PDB) with less than 95% sequence similarity to each other. The per residue false positive rate was found to be 3.2% on the set of proteins from the PDB, with a large fraction arising from short predictions of disorder that typically occur at the C- and N-termini.

Only 37 (0.5%) of the ordered structures were predicted to contain long (> 30 residue) regions of disorder. This value is likely to overestimate the false positive rate on this set, as many of the chains were crystallized as part of structural complexes and may be disordered prior to the formation of quaternary structure (see Figure 3.6). Other predicted regions of disorder occur in regions of the protein that form domain linkers (see Figure 3.7), which may be disordered in solution to allow structural uncoupling of two or more globular domains (Dyson and Wright, 2002). Some of the examples, such as the protein shown in Figure 3.8, which undergoes induced folding upon binding to the nuclear cap (Mazza et al., 2002), are interacting with a ligand and others are stabilized by binding to DNA as shown in Figure 3.9.

Four of the putative, false predictions of extensive disorder occurred in NMR structures where there were either virtually no secondary structure elements or where the predicted disordered regions were modelled by more than 20 isoforms, such as Figure 3.10. It therefore appears that the only certain false positives are the 5

Bound to ligand	6
Bound to protein	14
Bound to DNA	5
NMR (no structure/multiple models)	4
Ribosome complex	1
Domain linker	5
Surface of protein	5
Total	37

Table 3.3: The potential causes for false prediction of long disordered regions in sequences that have ordered structures recorded in the PDB. The results come from a set of 7169 proteins that are non-redundant at 95% sequence identity. In all cases where the protein is described as bound, the predicted disordered region is in contact with the ligand, DNA or other chain. The NMR models contained almost no visible helical or strand elements or had several models for the region of the protein predicted to be disordered. Only the 5 examples where disorder occurs on the surface of a globular domain are certain to be ordered in their native state.

(0.06%) cases, which occur on the surface of a globular domain without contact to another molecule. The potential causes for predicted regions of disorder being found in ordered structures are summarized in table 3.3.

This section has established a per residue false positive rate of 3.2% for the DISOPRED2 predictions of disorder in ordered structures, and more importantly, a per chain false positive rate of less than 0.1% for long regions of disorder. The following section presents the results of applying DISOPRED2 to complete archaea, eubacteria and eukaryote genomes.

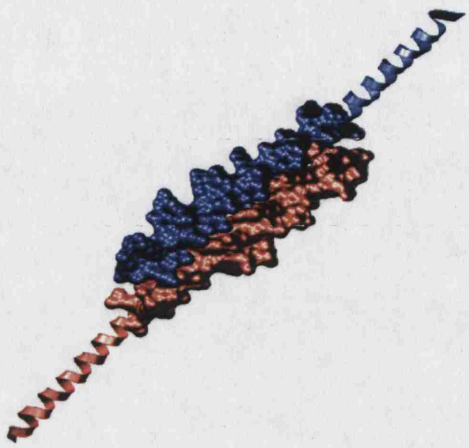

Figure 3.6: Structure of Bovine If1, the regulatory subunit of mitochondrial F-ATPase (1gmj) with predicted regions of disorder in the protein-protein interaction sites highlighted by the space-filling structures. All graphical representations of protein structure, except for Figure 3.10 which was generated using RasMol (Sayle and Milner-White, 1995), were generated using the VMD molecular visualization software (Humphrey et al., 1996).

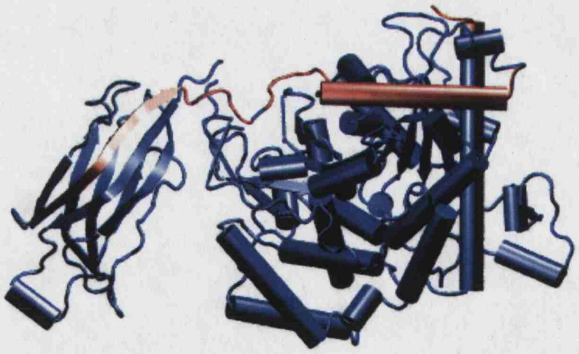

Figure 3.7: Structure of human cytosolic phospholipase (1cjy). The predicted regions of disorder are coloured in red. The disordered region overlaps the domain linker, along with a helix element in the domain on the right and a sheet in the domain on the left.

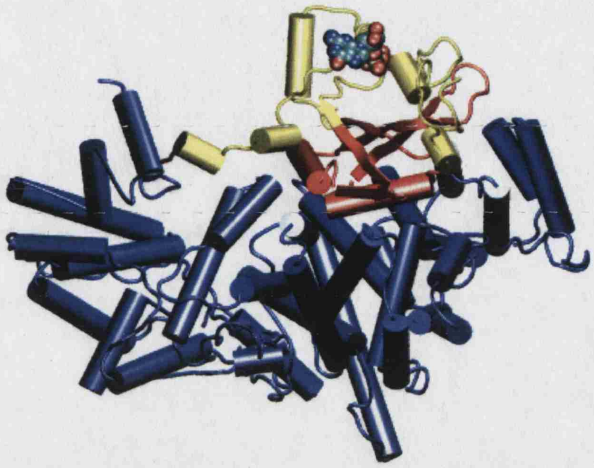

Figure 3.8: Structure of the Human nuclear cap-binding-complex (cbc) in a complex with cap analogue (M7Gpppg). The region of the protein that is predicted to be disordered is coloured in yellow, and is contained entirely within the cbc chain (coloured in red). The molecule in contact with the disordered region is the nucleotide GDP, and the cap analogue (M7Gpppg) is coloured in blue.

Figure 3.9: X-ray crystal structure of transcription factor (1gt0) from *Homo sapiens* bound to DNA. The protein structure is taken from a single chain with the apparent discontinuity caused by missing co-ordinates in the electron density map. The predicted disordered regions are coloured in yellow.

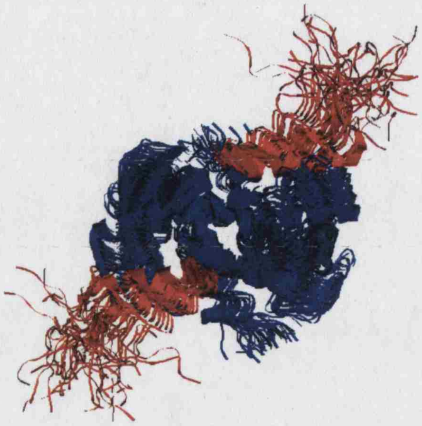

Figure 3.10: NMR structure of the C-terminal negative regulatory domain of *p53* in a complex with Ca^{2+}-bound S100B. Figure shows the ensemble of 40 model isoforms, and indicates significant dynamic flexibility in the regions of the protein that are predicted to be disordered. These putative disordered regions are coloured in red.

3.4.2 The Predicted Frequencies of Disorder in Complete Archaea, Eubacteria and Eukaryote Genomes

The protein sequences for 6 archaea, 13 eubacteria and 5 eukaryote genomes were downloaded from the National Center for Biotechnology Information (NCBI) ftp server. These sequences were first filtered using the sequence masking program *pfilt* to remove putative coiled-coil and transmembrane regions (Jones and Swindells, 2002). Regions with low sequence complexity and/or compositional bias were not masked out as these often represent disorder. The PSI-BLAST jobs, used to calculate the inputs to DISOPRED2, were distributed across a Linux beowulf cluster of Intel Pentium and AMD Athlon processors and two associated SunFire 880 servers running Solaris. Sequences were submitted to the Sun Grid Engine[3] scheduler using servlet technology.

[3]http://www.sun.com/gridware

Table 3.4 shows estimated disorder frequencies for 6 archaeal, 13 bacterial and 5 eukaryotic genomes, in addition to overall totals for each kingdom and predictions from a non-redundant set of resolved crystal structures in the PDB. An average of 2.0% of archaeal, 4.2% of eubacteria and 33.0% eukaryotic proteins are predicted to contain long regions of disorder. The frequency of disordered segments of length > 30 and > 50 residues was counted to allow comparison with an earlier study (Dunker et al., 2000), and gain a rough indication of the number of functional regions of native disorder.

Figure 3.11 shows the fraction of proteins in the Archaea, Eubacteria and Eukaryota that contain predicted disordered regions of length greater or equal to thresholds which vary from 0 to 100 residues. This plot shows that a larger proportion of eukaryotic proteins contain long disordered segments at all length thresholds than do proteins in the other two kingdoms of life. The cumulative frequency distribution of proteins from the three kingdoms of life as a function of disorder composition is shown in Figure 3.12. This plot shows that a larger proportion of eukaryotic proteomes are predicted to be disordered.

3.5 The DISOPRED2 Server

The very frequent occurrence of long disordered segments in eukaryotic proteomes suggests there are limitations on the number of structures that can be determined using X-ray crystallography because of the difficulties in purifying, crystallizing and resolving proteins that contain significant disorder. The DISOPRED2 server was therefore developed to provide a web interface that allows experimentalists to obtain predictions of the dynamically disordered regions in a protein sequence.

Kingdom *Organism*	Number of sequences	disorder frequency	length > 30	length > 50
Archaea *Aeropyrum pernix*	1841	4.7	2.1	0.5
Archaea *Archaeoglobus fulgidis*	2409	2.8	0.9	0.2
Archaea *Halobacterium sp.*	2442	6.2	5.0	1.9
Archaea *Methanococcus jannaschii*	1784	2.8	1.0	0.3
Archaea *Pyrococcus abyssi*	1769	3.0	1.4	0.7
Archaea *Thermoplasma volcanium*	1497	3.2	1.0	0.3
Bacteria *Agrobacterium tumefaciens*	5288	6.4	5.7	2.0
Bacteria *Aquifex aeolicus*	1557	3.3	1.9	0.4
Bacteria *Chlamydophila pneumoniae*	1111	6.2	4.8	2.3
Bacteria *Chlorobium tepidum*	2248	5.1	3.3	0.5
Bacteria *Escherichia coli*	4247	4.6	2.8	0.8
Bacteria *Haemophilus influenzae Rd*	1650	4.4	3.8	1.3
Bacteria *Mycobacterium tuberculosis*	3890	9.1	7.0	3.3
Bacteria *Neisseria meningitidis*	2020	5.7	4.5	1.7
Bacteria *Salmonella typhi*	4714	4.9	2.7	0.9
Bacteria *Staphylococcus aureus*	2632	6.2	4.5	2.2
Bacteria *Synechocystis species*	3140	5.4	4.7	1.8
Bacteria *Thermotoga maritima*	1857	3.3	1.8	0.6
Bacteria *Treponema pallidum*	1035	6.1	6.4	2.6
Eukaryota *Arabidopsis thaliana*	21482	16.8	33.8	19.0
Eukaryota *Caenorhabditis elegans*	20506	15.9	27.5	15.6
Eukaryota *Drosophila melanogaster*	13913	21.6	36.6	22.1
Eukaryota *Homo sapiens*	26385	21.6	35.2	21.9
Eukaryota *Saccharomyces cerevisiae*	6245	17.0	31.2	19.3
Archaea	11742	3.9	2.0	0.7
Bacteria	35389	5.7	4.2	1.6
Eukaryota	88531	18.9	33.0	19.6
PDB	7169	3.2	0.5	0.1

Table 3.4: **Estimated Disorder Frequencies.** The columns show the number of sequences, the percentage of residues predicted as being disordered and the percentage of chains with contiguous disordered segments of length greater than 30 and 50 residues, respectively.

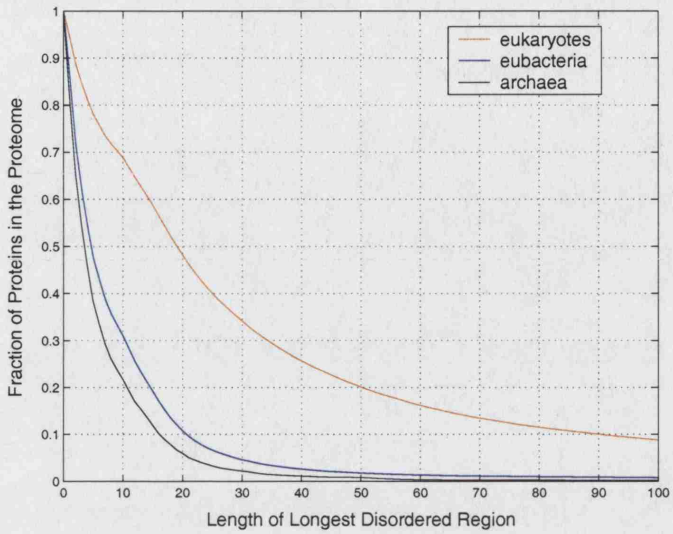

Figure 3.11: Proportion of proteins from the Archaea, Eubacteria and Eukaryota that contain disordered segments with lengths greater than variable threshold. For example, around 0.1 (10%) of eukaryote proteins are predicted to have a disordered segment of length greater than or equal to 90 residues.

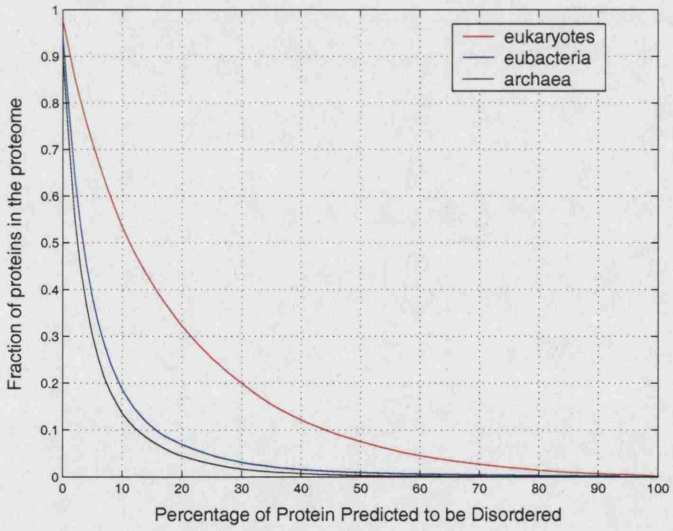

Figure 3.12: Fraction of proteins from the three kingdoms of life that have a predicted disorder composition greater than a threshold which varies between 0 and 100% of the total length. A factor of around 0.2 of all eukaryote proteins are predicted to be more than 30% disordered.

3.5.1 Server Design

A web form uses Java Servlets to submit user options and the amino acid sequence to a shell script which runs the initial PSI-BLAST search. The amino acid sequence is checked for the presence of non-letter characters, and other variables are cast to integers to provide security against malicious submissions. Error messages or confirmation that the server has received the request successfully are returned using Java Servlet Pages (JSP). After generation of the PSI-BLAST profile, the shell script runs a C++ executable to obtain the disorder predictions. This executable uses the primal weights in order to reduce the time complexity of the classifier.

The dual form of the decision function $f(\mathbf{x})$ for classifying an example vector $\mathbf{x}$, where N is the number of support vectors $\mathbf{x}_i$, is given by

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b = \sum_{i=1}^{N} \alpha_i \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) \rangle + b \tag{3.9}$$

which for the linear kernel implies

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \tag{3.10}$$

The weight vector $\mathbf{w} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i$ is therefore easily calculated from the Lagrange multipliers α_i and the support vectors.

The predictions are saved in a horizontal text format for returning to the user by e-mail with numerical outputs written to a plain text file that is used for generating disorder profile graphics.

3.5.2 Description of Server Use

Single letter amino acid sequences can be pasted into the DISOPRED2 server with the results delivered to the user by e-mail. The server makes several options available, including the return of hits and/or the alignments from the PSI-BLAST search. The web form also provides an option for setting the estimated false positive rate of the classifier. This allows the user to alter the precision and recall characteristics of the prediction, which are defined as

$$\text{precision} = \frac{TP}{TP+FP} \tag{3.11}$$

$$\text{recall} = \frac{TP}{TP+FN} \tag{3.12}$$

where TP is the number of disordered examples correctly classified. FP and FN are the numbers of false positive and false negative predictions of disorder. Receiver Operating Characteristic (ROC) curves and precision/recall tables are included in the help section. The results from DISOPRED2 are sent in plain text format along with hypertext links to PostScript, Portable Document Format (pdf) and JPEG images, which show plots of the sequence disorder profile (Figure 3.13). This allows the user to set arbitrary decision thresholds by visual inspection.

PSIPRED secondary structure predictions are also included to provide further structural information on the protein (Jones, 1999). PSIPRED predictions use similar inputs to DISOPRED2 and accurate predictions can be included with little computational overhead. If this option is checked, links to graphical representations of the predictions are provided using the PSIPREDView Java application (McGuffin et al., 2000). Figure 3.14 shows a screen shot of the DISOPRED2 home page and help section. Figure 3.15 shows a screen shot of the web form for submitting disorder predictions to the DISOPRED2 server.

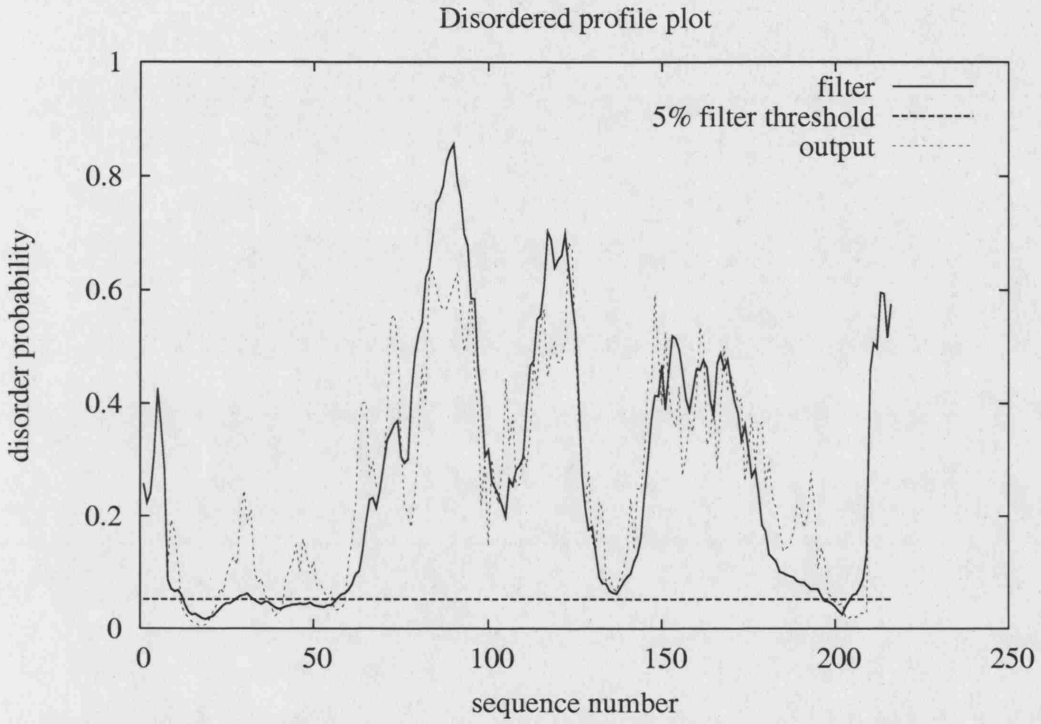

Figure 3.13: Example prediction for the intracellular loop of the membrane protein gliotactin from *Drosophila*. This sequence was classed as completely unstructured in the CASP5 experiment. The plot shows position in the sequence against probability of being disordered. In total, 157 of the 216 residues are classed as disordered at the default threshold. The horizontal line is the order/disorder threshold for the default false positive rate of 5%. The 'filter' curve represents the outputs from DISOPRED2 and the 'output' curve the outputs from DISOPREDsvm. The outputs from DISOPREDsvm are included to indicate shorter, low confidence predictions of disorder.

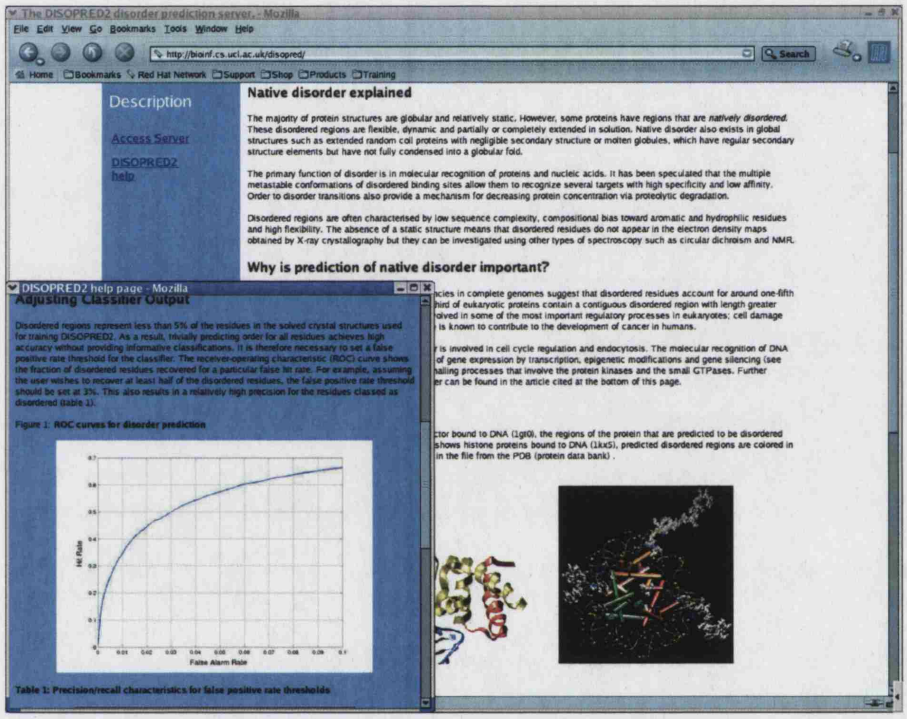

Figure 3.14: Screen shot of the home page and help section for the DISOPRED2 server. The home page includes a description of the prediction method, a brief explanation of the functional significance of native disorder and links to the help pages.

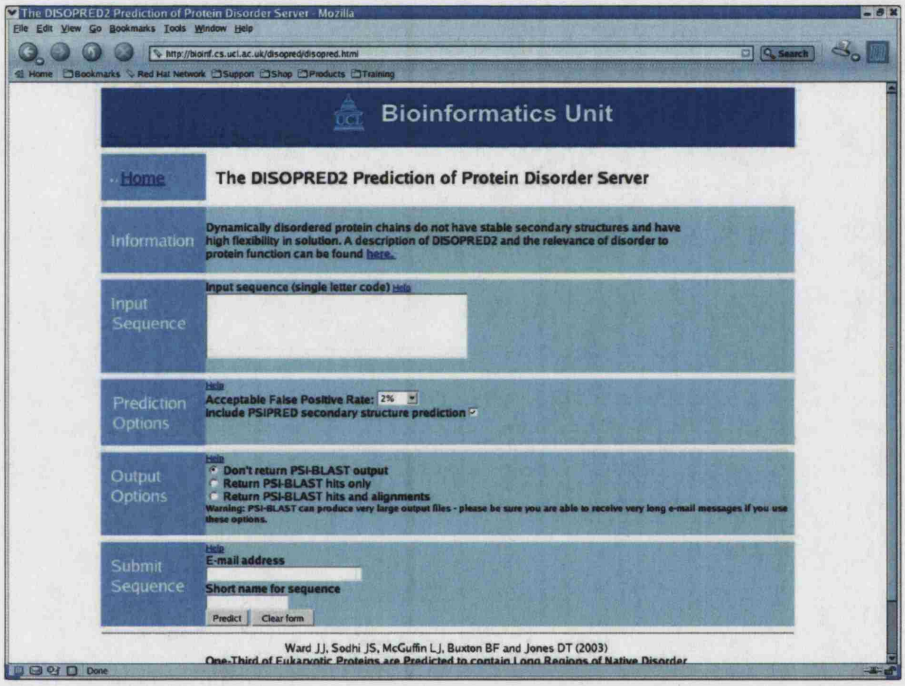

Figure 3.15: Web-page for the DISOPRED2 server. The form includes text boxes for inputting amino acid sequences and check boxes for various options.

The DISOPRED2 server was made available in April 2004, and had received 1015 submissions by the following June. An executable version of DISOPRED2 was made publicly available in July 2004. The executable version is currently being used in the MSGC[4] structural genomics project for both target selection and for identifying local regions of disorder that may be hindering crystallization of proteins that are predominantly ordered. An Applications Note advertising the DISOPRED2 server was published in September 2004 (Ward et al., 2004a) in the journal *Bioinformatics.*

3.6 Discussion

The difficulty in investigating dynamically flexible polypeptide sequences is the main reason for the relative paucity of experimental data on native disorder compared with globular structures. This difficulty also extends to the identification of disordered regions for the purposes of pattern recognition. The definitions of native disorder used by other authors (Wright and Dyson, 1999; Uversky et al., 2000; Dunker and Obradovic, 2001) are also fairly heterogeneous, as they apply to global structures such as collapsed molten globule proteins and extended random coil-like proteins, and to the localized disorder that can exist in flexible domain linkers and ligand binding sites.

In the training of DISOPRED2, residues with missing atomic co-ordinates are defined as disordered. Although this definition was also used in the CASP experiment, it is imperfect since missing residues can also arise as an artefact of the crystallization process such as rigid body wobble or crystal contacts. It is also possible that false prediction of order could be caused by the crystallized fragment being part of a structural complex or a multi-domain protein *in vitro.* However, this appears to be one of the simplest and most effective means of identifying disordered regions in

[4]Midwest Structural Genomics Consortium

the absence of further experimental characterization of the protein structure.

The development of DISOPRED2 has demonstrated that information from homologous sequences leads to a slight improvement in the prediction of native disorder. However, the improvement is not as great as that observed in predicting conventional secondary structure (Rost and Sander, 1993). This may occur as a result of disordered regions not being subject to the global constraints that are involved in the formation of globular proteins. If this is the case, then the local properties of the sequence are likely to contain all the necessary information for accurately predicting disorder. The slight improvement in using profiles rather than single sequences may also be explained by disorder being no more or less conserved than other parts of the protein (Liu et al., 2003). For example, it might be expected that disordered regions in molecular recognition sites would be under strong negative selection whereas those in domain linkers may not be conserved. These, and other similar conflicting signals in the conservation of disordered sequences, may reduce the effectiveness of sequence profiles.

There are likely to be several other reasons for the improved accuracy of the DISOPRED classifiers compared with the algorithms described in the methods section. The main difference from the other methods is that the DISOPRED methods are trained directly on protein sequence rather than measures of amino acid composition, sequence complexity (Vucetic et al., 2003; Romero et al., 2001) or biophysical properties such as mean hydrophobicity (Uversky et al., 2000). This may allow the classifier to recognize sequence motifs that have been shown to be associated with disorder such as Pro-X-Pro-X-Pro or Lys-X-X-Lys-X-Lys (Lise and Jones, 2004). The cascaded classifier, used in DISOPRED2, also improves accuracy by increasing the confidence in long predicted segments of disorder at the expense of shorter predictions. A small contribution to the improved accuracy may come from training set, which is taken exclusively from high resolution crystal structures, and does not

restrict the definition of disorder to long continuous regions.

A weakness of the analysis using structures from the PDB is that it is restricted to proteins that have been successfully crystallized, which do not constitute a random sample. The subset used to estimate error rates does not therefore contain members of structural families in the same proportions as those in the population, and there is also bias toward small, single domain proteins. However, the estimates are likely to be conservative because of the very low false positive rate and the likelihood of there being a significant number of disordered regions that are falsely predicted as ordered. It is possible, for example, that the under-representation of "unknown" annotations in the disordered set could be caused by DISOPRED2 failing to recognize the uncharacterized types of disorder that may exist in these proteins (see Chapter 4).

Although the most striking feature of Table 3.4 is the discrepancy between eukaryotes and prokaryotes, smaller differences are also observed between the archaea and the eubacteria. The scarcity of disordered regions in the thermophiles is perhaps caused by the strong evolutionary constraint on protein melting point in these organisms. Indeed, the only reference archaeal organism with an optimum growth temperature below $60^{o}C$ is the *Halobacterium* species, which is predicted to have a far larger proportion of long disordered segments than the other archaea. Amongst the eubacteria, the anomalously high prediction of disorder in *Mycobacterium tuberculosis* may be a result of its high $G + C$ content and a raised propensity toward the amino acids Gly, Pro and Arg (Cole et al., 1998).

Previous studies (Dunker et al., 2000; Vucetic et al., 2003) have found disorder to be ubiquitous in all three kingdoms of life with $\sim$ 60% of eukaryote, $\sim$ 28% of eubacterial and $\sim$ 36% of archaea proteins predicted to possess disordered regions longer than 40 residues. In this study, the comparatively low disorder over-prediction

rates (0.5% c.f. 17%) mean that large systematic differences in the error rates between organisms are less likely. Table 3.4 therefore provides convincing evidence for disorder being common in eukaryotes but much less so in prokaryotes. This confirms much of the experimental evidence to date, which has shown that dynamic flexibility of the protein structure is more often associated with eukaryotic protein function (Wright and Dyson, 1999).

There are several explanations for the lower occurrence of disorder in prokaryotes. Prokaryotes are subject to strong selective pressure on biochemical efficiency and do not have highly-regulated degradation pathways such as ubiquitination, so the cost of short protein lifetimes that arise from rapid proteolysis of disordered proteins is likely to be far greater. The absence of cell compartments also reduces the ability of prokaryotic cells to protect unfolded structures from degradation. The greater complexity of signal transduction, cell division and transcription in eukaryotes may also have acted to select positively for disordered structures because of their advantages in performing molecular recognition and other biochemical functions. The functions and evolutionary origin of disordered sequences are investigated in greater depth in the following chapter.

Chapter 4

Investigating the Functions of Disorder in *Saccharomyces cerevisae*

The clear disparity between the predicted disorder rates in eukaryote and prokaryote genomes, shown in Figure 4.1, could be explained by disorder being selected positively for various eukaryotic protein functions or that weaker purifying selection against short protein lifetimes in eukaryotes has led to an increase in redundant sequences. These two hypotheses are investigated in this chapter, which examines the functional relevance of putative, long regions of disorder in the budding yeast *Saccharomyces cerevisiae.*

It has been shown experimentally that disordered regions are involved in some DNA binding interactions (Weiss et al., 1990) and several other types of molecular recognition (Kriwacki et al., 1996). One speculated functional advantage of disordered binding sites is that their adoption of several transient structural conformations allows recognition of multiple targets with high specificity and low affinity (Dunker et al., 2002). The rapid turnover of proteins with unfolded portions also provides a means for altering protein concentration in response to cell cycle or extracellular stimuli (Nakayama et al., 2001). Transitions between the native unfolded state and a globular structure, induced by phosphorylation or some other type of interaction, also provide a mechanism for regulation of protein activity (Radhakrishnan et al., 1997). Finally, the flexibility of domain linkers is also a structural characteristic of many multi-domain proteins, and is functionally important for active sites that are formed between globular domains (Dyson and Wright, 2002).

This chapter begins by giving a brief summary of some of the experimental and computational studies that have provided evidence for the functionality of intrinsically disordered protein structures. This is followed by a description of the Gene Ontology (GO), which provides a universal language for describing the function of gene products (Gene Ontology Consortium, 2001), and the rationale for using the sequences from the budding yeast *Saccharomyces cerevisiae* for exploring the link between disorder and protein function in eukaryotic cells.

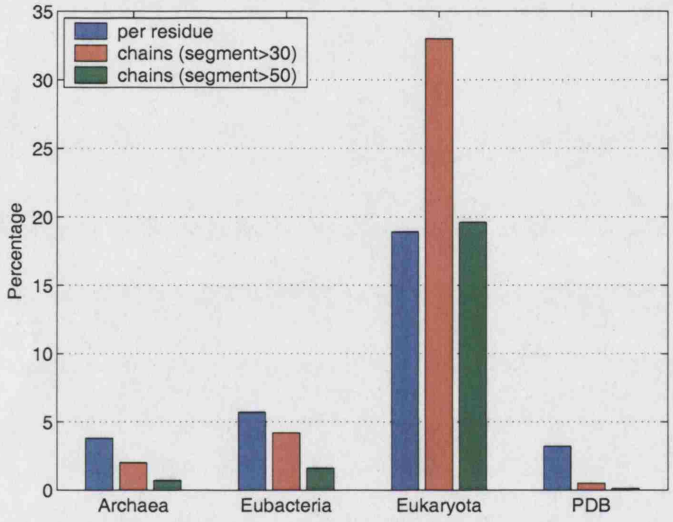

Figure 4.1: Predicted frequency of disorder in set of Archaean, Eubacterial and Eukaryotic genomes. The entry for the PDB is calculated solely from ordered structures, and therefore provides an estimate of the false positive rate. The legend refers to the per residue frequency of disorder and to the percentage of chains that contain a contiguous disordered segment longer than 30 and 50 residues.

A statistical analysis of the coincidence of boundaries between structural domains and predicted disordered segments is then presented using the ASTRAL domain cuts defined on the SCOP database (Chandonia et al., 2004; Murzin et al., 1995). This motivates the design of the sampling method for testing the significance of functions associated with long regions of predicted disorder. The relationship between putative disordered regions and specific Gene Ontology terms in the Saccharomyces Genome Database (Gene Ontology Consortium, 2000; Dwight et al., 2002) is then compared to previous genome-wide studies of native disorder.

The following section describes some experimental studies of native disorder in greater detail, in addition to results from the Dunker-Obradovic groups on the relationship between predicted disorder and protein function.

4.1 Experimental and Computational Studies of the Functions of Disorder

The importance of disorder to the biochemical function of many proteins has been discussed in several recent reviews (Wright and Dyson, 1999; Dunker and Obradovic, 2001; Tompa, 2002). These reviews are summarized in this section and supplemented with results from several recent experiments that provide further insight into the functions of disorder. A recent review of intrinsically unstructured proteins provided experimental examples of disorder taking part in DNA/RNA recognition, post-translation modifications, cell cycle regulation and membrane fusion/transport (Wright and Dyson, 1999). These various functions have been organized into five general categories by Tompa (2002). The five classes are:

1. **Entropic chains** In these proteins, the physical properties of disordered chains are essential to the particular function. Examples include the PEVK

domain[1] of elastic titin which acts as an entropic spring in muscle fibres, and the conformational freedom provided by the flexibility in some domain linkers.

2. **Effectors** The disordered binding sites within these proteins are involved in regulation *via* interactions with DNA or other proteins. Disordered binding sites are believed to allow these proteins to reversibly bind multiple targets with high specificity. Examples include the 4EBP translation initiation inhibitor and the transcription activator/repressor protein RAP1.

3. **Scavengers** These proteins are involved in the sequesterization of ions and small ligands. The increased surface area of proteins that lack globular structures allows binding of a larger number of small molecules. The caseins, which act as calcium-phosphate precipitation inhibitors in milk, contain disordered regions and have a binding site activity of $10^6 - 10^7 s^{-1}$. This rapid reaction rate is enhanced by an intermediate non-specific interaction between the disordered binding sites and the ligand (Tompa, 2002).

4. **Assemblers** This class of proteins are involved in the assembly, stabilization and regulation of macromolecular complexes such as the ribosome, cytoskeleton and transcription preinitiation complex. In assembly proteins, unstructured regions provide the flexibility in 'hinge' and 'pivot' regions that allow the mechanical motions required for assembling the complex from its component sub-units. Disorder is also involved in domain rearrangement and binding-induced folding between the proteins that form the complex. An example is the MAP2 microtubule-binding domain, which is active in the polymerization of tubulin dimers.

5. **Display Sites** Disordered regions are also involved in presentation sites, either for post-translational modifications or degradation by proteases. It is

[1]Rich in the amino acids proline, glutamic acid, valine and lysine.

known that disordered regions are very sensitive to protease degradation *in vitro*, which may explain the short lifetimes of proteins that contain PEST sequences[2], and may be a useful property for proteins involved in signalling cascades. Disordered regions are also present in the sequences flanking many serine and threonine protein phosphorylation sites (Iakoucheva et al., 2004).

The final four categories, in the list above, are processes involving molecular recognition, and this appears to be the principal role of disorder in eukaryotes.

The review by Wright and Dyson (1999) stated that the mechanism for the increased specificity of disordered binding sites is the large loss of conformational entropy resulting from the formation of a binding-induced structure. If, for example, the interaction between a disordered binding site and a particular DNA sequence motif is to be favourable thermodynamically, the increase in the entropy contribution, $-T\Delta S$, to the Gibbs free energy, ΔG, must be compensated by a decrease in the enthalpy term, ΔH, which Wright and Dyson (1999) argued is only sufficient for the cognate DNA sequence

$$\Delta G = \Delta H - T\Delta S \tag{4.1}$$

where T is the temperature. This allows recognition of a specific DNA sequence motif to occur reversibly because of the low affinity of the interaction but with very high specificity. It has also been shown that a loss of conformational entropy is involved in regulating the protein-protein interaction between the phosphorylated form of the kinase-inducible domain pKID of CREB with the KIX domain of CBP (Radhakrishnan et al., 1997).

[2]Rich in the amino acids proline, glutamic acid, serine and threonine. Figure 3.4 indicates that three of these residues have a high propensity for forming disordered structures.

A recent NMR study of the specific and non-specific interactions between the lactose repressor, *lac*, and DNA provides an example of flexible protein structures mediating transcription factor activity (Kalodimos et al., 2004), and suggests another functional advantage of disorder in DNA-binding domains. The *lac* repressor binds non-specifically to DNA as a result of electrostatic interactions between flexible, positively charged side-chains and DNA. The relatively low affinity of this interaction allows *lac* to undergo 1-dimensional diffusion along the length of the DNA molecule. This low affinity, non-specific binding step allows the recognition of specific DNA sequence motifs to occur at a rate that is a factor of $10^2 - 10^3$ faster than the rate predicted for a 3D diffusion-controlled reaction. Upon binding to a cognate DNA sequence, a disordered hinge region forms a helix that greatly increases the binding affinity and mediates repressor activity (see Figure 4.2). The increase in association rate, which is facilitated by a disorder-order transition, is another functional advantage of native disorder in DNA-binding proteins.

The various transient structural conformations of intrinsically disordered proteins also provide binding diversity. This is particularly important for the hub nodes in signalling networks, which regulate the activity of multiple targets. The most famous example is tumour suppressor protein *p53*, which regulates transcription in the nucleus and is activated by interactions with other proteins and post-translational modifications in response to cellular stress. *p53* can also induce apoptosis by several different mechanisms including alteration of gene expression and translocation to the mitochondria (Haupt et al., 2003). Under normal conditions *p53* is short-lived, as a result of its unstructured domains, but persists longer when its proteolysis is inhibited as part of the cellular response to stress.

Dunker and Obradovic have also published results on the functions of disordered proteins (Dunker et al., 2002), and the tendency of cell-signalling and cancer-associated protein sequences to be predicted to contain long regions of disorder

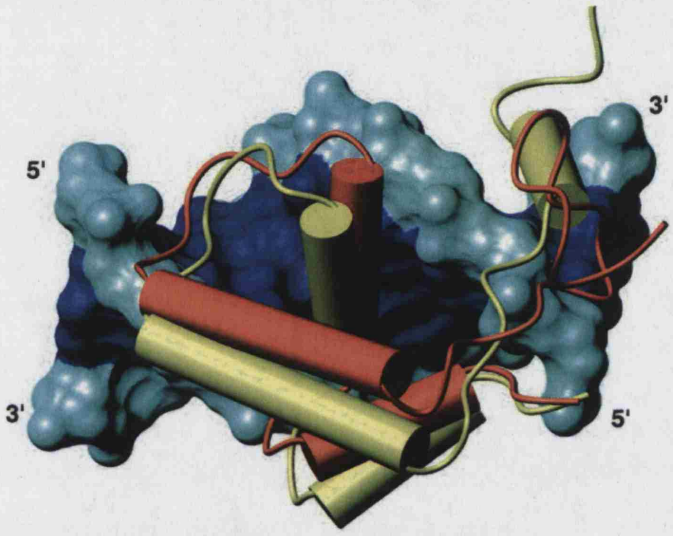

Figure 4.2: Specific (yellow) and non-specific (red) binding modes of the *lac* repressor with DNA. The protein in the non-specific complex is rotated 25^o, relative to the DNA, compared to the protein in the specific complex. The helix interacting with the minor groove in the specific complex is also unstructured in the nonspecific complex. Figure reproduced from Kalodimos et al. (2004).

(Iakoucheva et al., 2002), in addition to work on the properties of disordered sequences (Romero et al., 2001) and the estimated frequency of disorder in complete genomes (Dunker et al., 2000; Vucetic et al., 2003).

The most recent study by the Dunker-Obradovic group involved the application of the VLXT method (Iakoucheva et al., 2002) to the entire SWISS-PROT sequence database (Bairoch and Apweiler, 2000). The false positive rate of the VLXT classifier was again estimated using a set of ordered PDB structures with less than 25% sequence identity. The per chain error rate for long regions of disorder was found to be 13±4% on this set. VLXT was then applied to all the sequences in SWISS-PROT, and used to calculate the disorder composition and the probability of containing a long (> 30) segment for various functional classes, which were defined using SWISS-PROT keywords.

The principal finding was that the human sequences annotated with the keywords 'anti-oncogene', 'oncogene', 'proto-oncogene' or 'tumor' (collectively referred to as human cancer-associated proteins) were significantly enriched in long predicted regions of disorder ($79 \pm 5\%$) compared with eukaryotic proteins in SWISS-PROT ($47 \pm 4\%$). Iakoucheva et al. (2002) also found that the keywords 'regulation', 'cytoskeleton' and 'inhibitor' were associated with sequences that had a high predicted disorder composition. Sequences with 'kinase', 'metabolism', and 'Glyco-protein coupled receptor' keyword annotations had the lowest estimated disorder composition. Although this represented a large and comprehensive study of the functions of long regions of putative disorder, there were several deficiencies, which are addressed in this chapter.

The first improvement is that the DISOPRED2 prediction method with its relatively high accuracy and low per chain error rate is used instead of VLXT. The analysis is also carried out on the Saccharomyces Genome Database, which has the

advantage that it is representative of a typical eukaryote proteome, contains fewer sequencing and gene identification errors than sequences from other eukaryotes in SWISS-PROT or other large sequence databases, and is annotated using the Gene Ontology. Iakoucheva et al. (2002) also recognized that longer protein sequences have a greater probability of incorporating a disordered segment, and that this bias is reflected in some of the functions that they found to correspond with disorder. For example, the proteins annotated with the 'ribosome' keyword were found to have a lower probability of incorporating a long, putative, disordered region than cytoskeletal proteins. However, this could be caused by ribosomal proteins, with an average length of 187 residues, being much shorter than cytoskeletal proteins, which have a mean length of 1044 amino acids.

In Section 4.5, the associations between particular GO terms and long regions of disorder are tested using an improved sampling method that is not affected by the lengths of the proteins within a particular functional class. The following section describes the structure of the Gene Ontology and some of the the advantages of using GO for describing the function of gene products.

4.2 Describing the Function of Gene Products: The Gene Ontology

The previous section described the functions of disorder in the traditional manner; natural language, which has the advantage that it can be used to describe the properties of biological processes very precisely. However, the main deficiency of natural language as a means for representing knowledge is that it is difficult to extract and summarize this information automatically. The explosion of data from the genome sequencing projects and other organism-wide experiments has necessitated the de-

velopment of systematic and coherent strategies for describing biological phenomena. These schemes, and in particular *ontologies*, aim to provide a global representation of specific biological systems, on scales that range from the description of small molecules to entire ecosystems.

A particular problem in the post-genomic age has been the development of universal descriptions for the functions carried out by gene products. The Gene Ontology (GO) is the first attempt at developing a vocabulary for describing these roles across many scales and organisms. The early functional annotations were developed for the large sequence databases (Bairoch and Apweiler, 2000) and were based on keywords that allowed the user to perform simple queries. However, the absence of a systematic means for assigning these keywords and relating them to each other caused a reduction in the efficacy of the searches and precluded more sophisticated queries and cross-species comparisons. Other schemes such as the Enzyme Commission numbers (Webb, 1992) are limited to a subset of possible protein functions.

The GO consortium was formed to develop a universal vocabulary for describing gene products from several eukaryotic model organisms, originally including the yeast *Saccharomyces cerevisiae*, the fruitfly *Drosophila melanogaster* and the mouse *Mus musculus*. Since then, the number of genomes annotated using GO has grown to around twenty and includes the nematode worm *Caenorhabditis elegans*, the vascular plant *Arabidopsis thaliana*, *Homo sapiens* and several bacteria (Gene Ontology Consortium, 2001). The gene ontology therefore provides semantics that encapsulate biological information in a wide variety of organisms.

GO consists of three separate ontologies dedicated to molecular function, biological process and cellular component. Molecular function describes the biochemical reaction in which the gene product (protein or RNA) participates. Broad functional classes might be 'enzyme' or 'transporter' with more specific examples being

'Ca^{2+}/Calmodulin-dependent protein kinase II' or 'connexin'. The biological process refers to the gene products' participation in cellular or physiological processes, such as 'DNA-metabolism' or 'endocytosis'. Finally, the cellular component describes the location where the gene product is biologically active. These can include cellular compartments such as 'nucleus' or structural complexes such as 'histone deacetylase complex'.

The terms in GO are arranged in a directed acyclic graph (DAG) structure. Each term is represented by a node in the graph that is the child of one or more parent terms[3]. The child-parent relationships are either of the 'is a' type which states that the child term is a sub-category of the parent, or the 'part of' type which indicates that the child is a component of the parent (e.g. nuclear pore is part of the nuclear membrane). Every GO term is assigned a unique identifier although they are permitted to have several synonyms to include the nomenclature of the various GO contributors and other international standards.

The advantage of GO's DAG structure is that it is sufficiently general to capture semantic relationships that are formed between a child and multiple parents that may be unrelated to each other. However, this complicates matters from a classification perspective as each term does not represent an equivalence class at a particular level of granularity, in contrast to hierarchical schemes such as enzyme classification (EC) number or the MIPs functional assignments (Mewes et al., 2004).

The deviation from a hierarchical structure is most pronounced in the molecular function ontology where, for example, DNA helicase is a child of both the enzyme and nucleic acid binding terms. Another difficulty that applies particularly to molecular function and cellular component is that a gene product may be described by more than one term in the ontology, corresponding to multiple biochemical functions

[3]Apart from the root node; the Gene Ontology.

and cellular locations. These complications make automatic function assignment technically difficult and the most prominent attempts at annotating function have been limited to generating pair-wise relationships between protein sequences that show a greater overlap of SWISS-PROT keywords than would be expected under a random assignment model (Marcotte et al., 1999b).

Other analyses of protein function have been based on the enzyme commission (EC) number (Todd et al., 2001), which is a hierarchical classification scheme for the biochemical reactions catalyzed by natural enzymes. The EC number is made up of four digits with the first representing the broad enzyme class, such as hydrolase, and subsequent digits more specific information on the type of reaction such as the type of bond and substrate. These studies are useful for investigating the link between structure and function, since enzymes represent a large proportion of the solved structures with known function in the PDB (Hegyi and Gerstein, 1999).

4.3 *Saccharomyces cerevisiae* as a Model for Eukaryotic cells

There are several reasons for the budding yeast *Saccharomyces cerevisiae* being one of the most popular experimental model organisms. *Saccharomyces* most important attribute is that it is a unicellular eukaryote with nuclear DNA and a similar complement of organelles to the metazoa. Yeast exists in its native state in cell culture and has other useful properties such as rapid growth, dispersed cells and a versatile DNA transformation system (Sherman, 1997). Yeast also shares some of the characteristics of prokaryote cells that have led to the rapid developments in bacterial molecular genetics.

Although the yeast genome contains more than three times the DNA of *E. coli*,

it is relatively compact by the standards of other eukaryotes, and this is one of the reasons for *S. cerevisiae* being the first eukaryotic genome to be sequenced (Goffeau et al., 1996). Determining the protein sequences that arise from eukaryote genomes is not trivial because of the difficulties in identifying translation-initiation signals and effects such as splice variants, polyadenylation and RNA editing, which result in multiple mRNA species being transcribed from a single DNA sequence. Despite the relative simplicity of RNA processing in yeast, the set of translated ORFs (Open Reading Frames) has been revised several times, with the most recent update of the translated ORFs being carried out using a comparison with the complete genome sequences of several close relatives (Kellis et al., 2003). It is therefore certain that the "complete" genomes of higher eukaryotes such as *Homo sapiens* will be subject to even greater revision over the next decade.

From a bioinformatics perspective, the advantages of investigating the yeast genome is that it is close to full completion and therefore contains fewer sequencing and gene identification errors than other eukaryotes (Phizicky et al., 2003). A large volume of data has also been produced by the experimental community engaged in the study of *Saccharomyces.* These include experiments on the scale of the entire organism such as 2-hybrid screens and pull-down assays for detecting protein-protein interactions (Gavin et al., 2002; Uetz et al., 2000), systematic gene 'knock-outs' (Winzeler, 1999) and microarray experiments. This wealth of experimental data has lead to the creation of the Saccharomyces Genome Database (SGD), which has a comprehensive set of experimentally-determined functional annotations along with references to the relevant literature (Dwight et al., 2002). The developers of the SGD were also involved in the creation of the Gene Ontology for describing the function of gene products, and GO annotations are used throughout (Gene Ontology Consortium, 2001).

4.4 Analysis of the Occurrence of Native Disorder in Domain Linker Regions

The prediction of domain linkers is assuming greater importance in structural bioinformatics, as identification of which parts of the sequence form globular domains is likely to represent an important first step in predicting the full quaternary structure of multi-domain proteins. This section investigates whether domain cuts have a propensity for being predicted as disordered by DISOPRED2. The analysis of domain linkers was carried out on PDB structures (Berman et al., 2000) with domain cuts recorded in the ASTRAL SCOP (version 1.65) database (Chandonia et al., 2004; Murzin et al., 1995). This was reduced to a non-redundant set using a threshold of 90% sequence identity between amino acid sequences.

The ASTRAL domain cuts, given in terms of the ATOM residue positions, were then mapped to the correct amino acid in the SEQRES records. Sequences where the constituent domains did not form a single, contiguous portion of the sequence were then removed from this set. This could be caused by residues being missed from the ATOM records (disordered) or a globular domain being discontinuous in the amino acid sequence. This procedure resulted in a dataset of 6630 sequences containing 635,310 residues and 861 domain cuts.

DISOPRED2 was applied to this set of sequences by running a three-iteration PSI-BLAST search to generate a sequence profile, and using this profile to classify the order/disorder states of each amino acid (see previous chapter). The 10 residues at the N- and C-termini were removed from this analysis, as these regions cannot contain a boundary between globular domains[4] and are often disordered. This resulted in a total of 502,710 internal residues for investigating the utility of disorder

[4]The shortest domains recorded in ASTRAL are 28 residues in length.

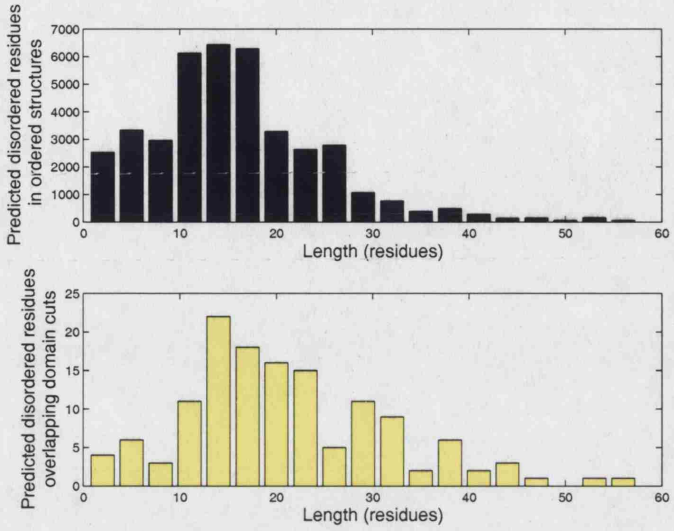

Figure 4.3: The upper plot shows the length distribution of the segments containing predicted disordered residues in the PDB. The lower plot shows the length distribution of the predicted disordered segments that coincide with ASTRAL domain cuts. The false positive rate threshold is 7% and the bins used to generate the plot are over three-residue increments.

predictions for identifying domain boundaries.

Figure 4.3 shows the length distributions of residues within disorder predictions that coincide with domain linkers and the 6630 sequences from the PDB overall at a false positive rate threshold of 7% (similar distributions are observed for other thresholds). This indicates that residues within longer regions of predicted disorder (> 10 residues) are more likely to coincide with a domain boundary than residues within shorter regions. Figure 4.4 shows a plot of the overall predicted disorder composition for the structures from the PDB against the disorder composition of the domain cuts, as defined by the ASTRAL database. The various curves were generated by removing predictions for disordered segments with lengths less than a threshold.

Figure 4.4 appears to show that disorder is over-represented in domain linker

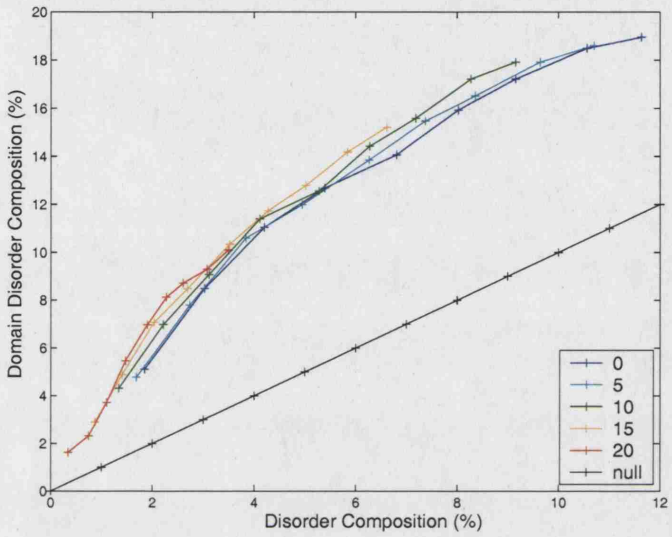

Figure 4.4: Relationship between the overall predicted disorder composition of a non-redundant set of ordered protein structures and the proportion of ASTRAL domain cuts that are predicted to be disordered. The coloured lines represent results of DISOPRED2 with predictions for segments with lengths below the threshold removed. The legend shows curves generated by removing disorder predictions that were less than 5, 10, 15 and 20 residues in length (in addition to 0 for not thresholding the length of the prediction). The points on these curves are generated by varying the bias of the DISOPRED2 decision function from a false positive rate threshold of 2% to 10%. The black line shows the expected proportion of domain cuts that are disordered under the null model (disorder occurs in domain cuts with the same frequency as in PDB structures).

regions. This hypothesis can be tested by assuming that each domain cut is a *Bernoulli trial* from the binary order/disorder distribution. The null distribution for successes (sampling a disordered residue) is therefore binomial and can be used to calculate a p-value for sampling more than n successes from N trials (domain cuts).

The probability of obtaining exactly n successes from N trials is

$$P_p(n|N) = \binom{N}{n} p^n (1-p)^{N-n} = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \tag{4.2}$$

where p is the probability of success (sampling a disordered residue). This is often approximated by a Gaussian distribution for large N and $p \simeq 0.5$. The probability, P, of obtaining more than n successes can be calculated exactly using the formula (Weisstein., 2004)

$$P = \sum_{k=n+1}^{N} \binom{N}{k} p^k (1-p)^{N-k} = I_p(n+1, N-n) \tag{4.3}$$

where

$$I_x(a,b) \equiv \frac{B(x;a,b)}{B(a,b)} \tag{4.4}$$

where $B(a,b)$ is the beta function $B(x;a,b)$ the incomplete function as given by

$$B(a,b) \equiv \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{(a-1)!(b-1)!}{(a+b-1)!} \tag{4.5}$$

$$B(x;a,b) \equiv \int_0^x u^{a-1}(1-u)^{b-1} du \tag{4.6}$$

The removal of all disorder predictions with lengths less than 10 residues, at a

false positive rate threshold of 9%, results in prediction of 33,209 disordered residues in the PDB[5], giving a value for p of 0.066. At these parameter settings, 148 (17.2%) of the total of 861 domain cuts are predicted as disordered. The p-value, calculated using Equation 4.3, for a larger number of domain cuts being disordered under the null model is 1.3×10^{-26}. This allows rejection of the null hypothesis, and indicates that domain boundaries have a propensity toward being predicted as disordered.

4.5 Investigating the Functions of Protein Disorder in *Saccharomyces cerevisiae*

The previous section suggests that the linkers between many globular domains are disordered. Although, this is a structural property of many multi-domain proteins and is involved indirectly in protein function, the aim of this analysis of predicted disordered regions is to determine which molecular functions rely directly on dynamic flexibility of the protein structure. In this section, functionally active disordered regions are identified with contiguous segments of the sequence longer than 30 residues that have predictions for disorder above the 2% false positive rate cut-off. This definition is imperfect, as these predicted regions of disorder may not be involved directly in the function of the protein, and disordered active sites with lengths shorter than the threshold are likely to exist. However, the estimated disorder over-prediction rate is extremely low at these thresholds ($< 0.1\%$, as shown in the previous chapter), so allowing only very confident predictions of disorder to be considered.

The proteins in the subset of the Saccharomyces Genome Database (SGD) predicted to contain long regions of disorder were longer on average than the population

[5]Excluding the amino acids at the domain cut.

(704.6 compared to 497.1 residues), which arises because large, multi-domain proteins have more linker regions and a higher probability of incorporating a disordered domain if these are distributed uniformly across the proteome. The sampling method accounts for this by constructing a null model where the disordered segments with lengths greater than 30 residues are distributed randomly across the length of the proteome. This is achieved by mapping each long disordered (> 30 residue) segment to the GO annotations attached to its parent protein. The frequency with which each GO term occurred is then compared to its frequency of occurrence in simulations (Efron and Tibshirani, 1993). The random model corresponds to a null hypothesis whereby each protein's probability of containing a long disordered segment is proportional to its length. A large number of replicates were then used to provide confidence estimates, under the null model, for GO terms that were over- or under-represented in the set of disorder predictions.

4.5.1 System and Data Sets

The analysis of the annotations that coincide with predicted disorder was carried out using the July 2003 release of the SGD. The SGD deals with the uncertainty of gene prediction by dividing the yeast ORFs into two subsets. The first set of 'verified' ORFs have been shown experimentally to be translated into proteins. The second set contains 'dubious' ORFs, which have a similar DNA structure to genes but are not known to be translated in significant numbers (Graur and Li, 2000). Only the translated ORFs, which include 2337 unique GO terms attached to 5889 proteins, are used in this section. At the 2% false positive rate threshold, 17.1% of the residues are predicted as disordered and 34.1% and 20.9% of chains are predicted to contain disordered segments longer than 30 and 50 residues, respectively. These frequencies differ slightly from those in Table 3.4, which were generated from the

NCBI curated version of the *Saccharomyces cerevisiae* proteome.

The majority of the GO annotations in the SGD are assigned by human experts, using all of the information available in the literature. However, some GO annotations are generated by automatically mapping descriptions from other schemes, such as Enzyme Commission (EC) numbers or SWISS-PROT keywords, to the appropriate GO term (Camon et al., 2003a). These 'electronic annotations' refer to a 'likely' or 'related' annotation to the correct GO term (Camon et al., 2003b), and were therefore excluded from this study to ensure reliability.

The SGD annotates each protein with the most specific terms available in the Gene Ontology. However, the hierarchical structure of GO means that all ancestral terms also provide a valid description in a more general sense. For example, the term "cell cycle" is *part of* "cell proliferation", which *is a* process of "cell growth and maintenance". Including the ancestral nodes in GO's directed, acyclic graph hierarchy expands the number of unique terms to 3299. In the graph of terms, it is possible for an ancestral node to be reached by more than one path, however in this study all of the ancestors of a particular term are counted only once as a function associated with the gene product.

A diagram of the sampling method, used to determine the functional significance of disorder, is shown in Figures 4.6 and 4.5. Each predicted, long disordered segment was first mapped to the GO annotations for the protein in which it occurred. The number of times each GO term was found to coincide with a disordered region was then counted across the entire set of disordered predictions. In each random simulation, segments with identical lengths to the disordered predictions were randomly distributed across the yeast proteome with the constraint that segments could not cross the boundaries separating each protein. The probability of a shuffled segment occurring within the span of each protein is therefore proportional to its length.

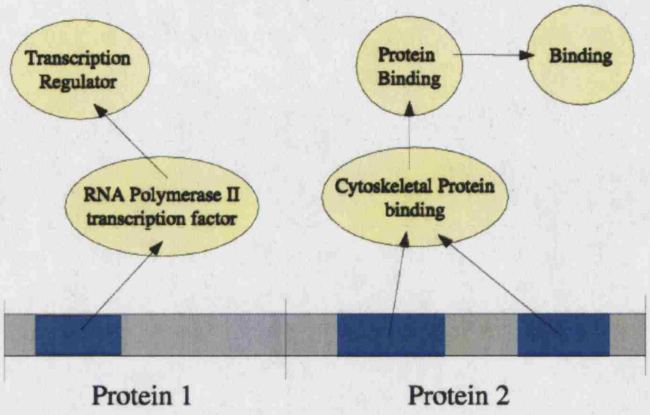

Figure 4.5: Schematic representation of long, predicted regions of disorder being mapped to the gene ontology annotations used to describe the sequence in which they occur.

The number of times each GO term occurred in 10,000 simulations was then used to obtain p-values for the disorder predictions under the null model.

4.5.2 Results and Corrections for Multiple Hypothesis Tests

GO terms that obtained a p-value lower than 0.02 and that describe more than 30 protein annotations[6] are shown in Figures 4.7, 4.8 and 4.9, which divide the results into the three separate ontologies representing molecular function, biological process and cellular component. The GO terms are ordered by the normalized differences between the terms' mean frequency of occurrence in the random samples and in the set of disordered predictions. The normalization factor is the standard deviation of the random sampling experiments.

Several GO terms that are used to describe similar functions and/or highly over-

[6] A two-sided p-value of 0.02 corresponds to fewer than 100 out of the 10,000 resamplings receiving a more extreme Z-score. Full results can be found at http://bioinf.cs.ucl.ac.uk/disopred/suppInfo.html

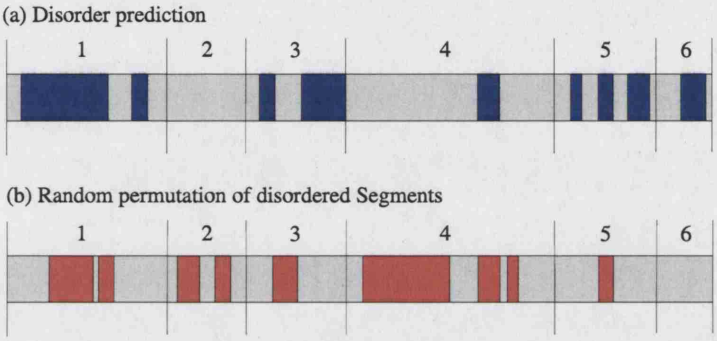

Figure 4.6: The two diagrams represent the yeast proteome with the six proteins separated by vertical lines. (a) DISOPRED2 is applied to the proteome, and a count is made of the number of times each GO term coincides with a putative, long disordered segment (coloured in blue). (b) Segments with the same length distribution are shuffled randomly across the proteome (coloured in red). 10,000 replicates are then used to calculate a null distribution for each GO term.

lapping sets of yeast proteins are omitted from the figures. For example, kinase, protein kinase, and protein serine/threonine kinase activity are all over-represented in proteins that contain long disordered segments. Of these nested sets of proteins, only protein kinase activity is included in Figure 4.7, as this term received the highest of the three Z-scores. This approach is also used to resolve the occurrence of similarly related terms in Figures 4.7-4.9.

It should be pointed out that, since a large number of hypotheses are being tested simultaneously, there is a high probability that some of the terms will be falsely evaluated as significant (type I errors). For example, if hypothesis tests are carried out on one-hundred independent physiological measurements on a large sample of diseased and normal patients, then 5 measurements would be expected to have p-value < 0.05 under the null model, even if there are no differences between the two underlying distributions. This can be controlled by introducing a Bonferroni adjustment to the p-value thresholds α' used to define statistical significance of each independent hypothesis.

$$1 - \alpha = (1 - \alpha')^k \tag{4.7}$$

where k is the number of hypotheses and α is the p-value of the global null hypothesis (all independent hypotheses are null). For small α this is approximated by

$$1 - \alpha = 1 - k\alpha' \tag{4.8}$$

$$\alpha' = \frac{\alpha}{k} \tag{4.9}$$

Although Bonferroni corrections decrease the potential for type I errors, they also reduce the *power* of the hypothesis test and increase the possibility of the null

model being accepted falsely (Perneger, 1998). These type II errors are arguably less acceptable in this type of study, which is attempting to demonstrate a useful application of the prediction method and to suggest potential directions for further experimental research.

A compromise was achieved by performing a Bonferroni adjustment for the sets of proteins annotating large numbers of yeast proteins. Terms that are statistically significant after the Bonferroni adjustment could then be considered confidently ($p < 0.01$) not to have occurred under the null model. Terms that were not statistically significant after the Bonferroni correction or that contained fewer proteins than the size threshold are also shown in Figures 4.7-4.9, as these may indicate functions that involve disorder, but with the proviso that these terms have a greater likelihood of being type I errors.

The Bonferroni corrections were carried out for each ontology separately. The number of independent hypotheses, k, was counted as the number of GO terms from the molecular function and cellular component ontologies that annotate more than 100 proteins in yeast. A threshold of 200 was used for biological process, as this ontology contains more terms annotating a large number of yeast proteins than the other two ontologies. The parameter, k, for adjusting the significance thresholds was therefore 37 for molecular function, 32 for cellular component and 45 for biological process. These values for k are conservative, as each GO term does not represent an independent hypothesis. This occurs because there is significant overlap in the sets of proteins annotated to particular GO terms. This tendency for GO terms to co-occur in the descriptions of yeast proteins arises as a result of functional relationships between terms, and particularly from the Gene Ontology's DAG structure (e.g. parent-child relationships such as between 'transcription factor' and 'DNA binding' activity).

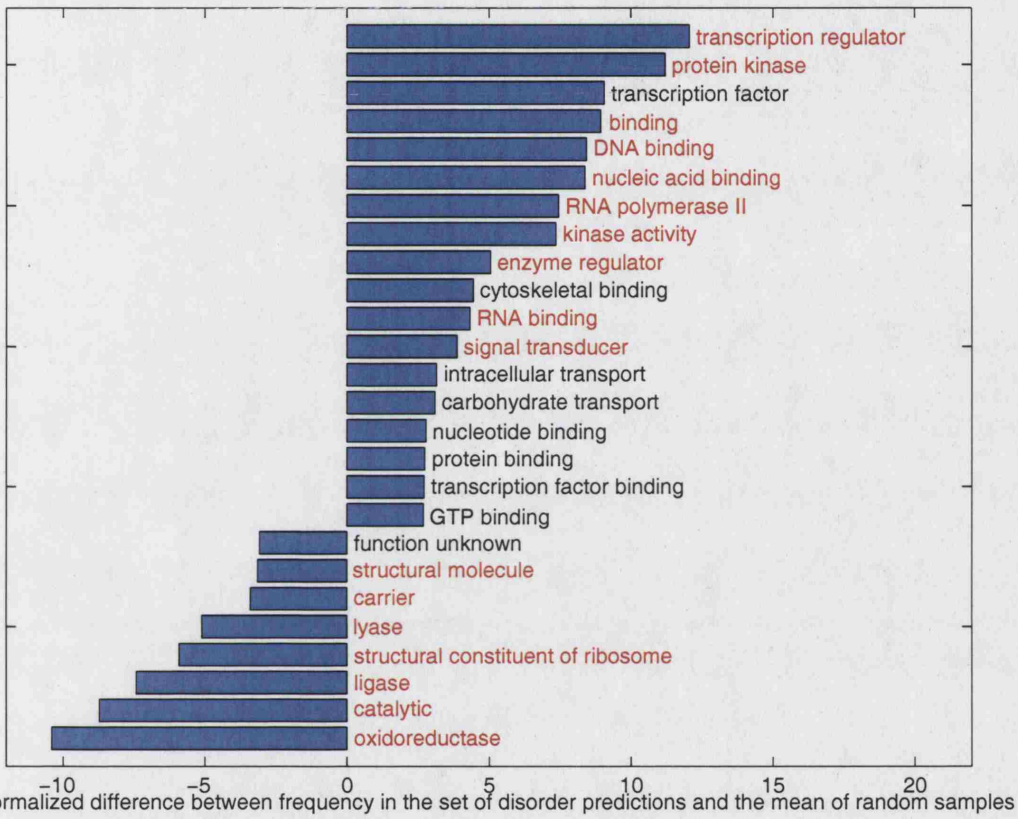

Figure 4.7: GO terms from the molecular function ontology that are significantly over- or under-represented in the set of proteins predicted to contain long regions of disorder. The terms are ordered by the normalized differences between the terms' mean frequency of occurrence in the random samples and in the set of disordered predictions. The normalization factor is the standard deviation of the random sampling experiments. The GO terms with names coloured in red annotate more than 100 proteins in yeast and have a p-value < 0.01 after the Bonferroni correction ($k = 37$).

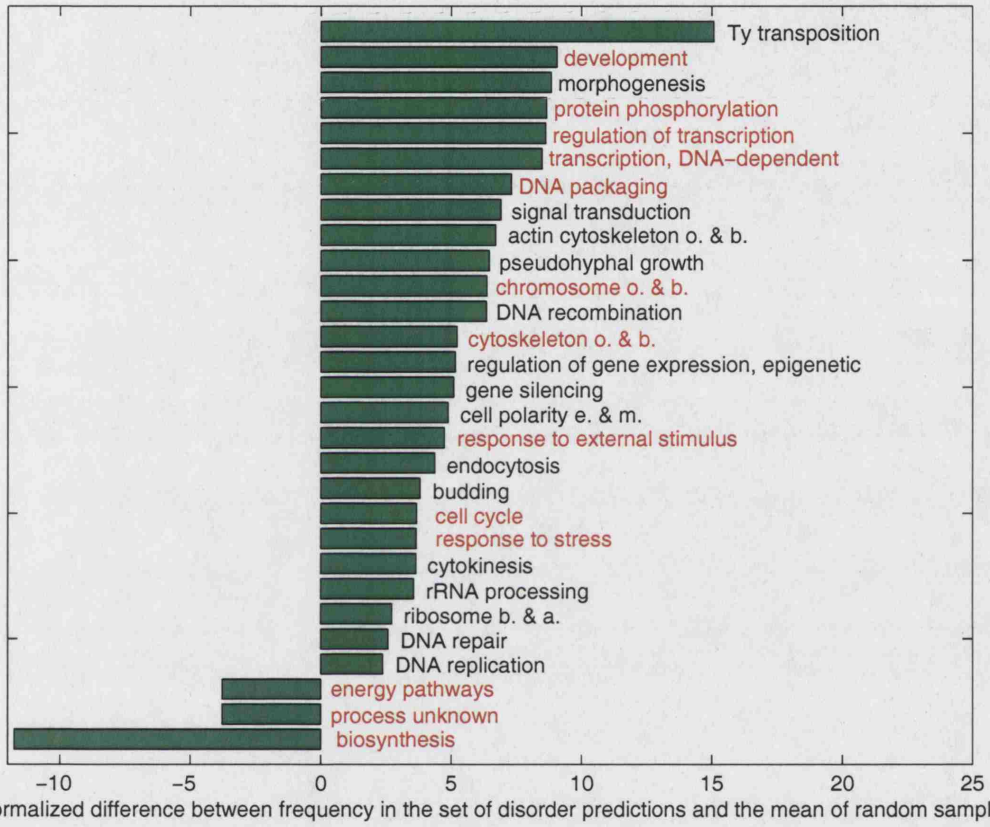

Figure 4.8: GO terms from the biological process ontology that are significantly over- or under-represented in the set of disordered predictions. The abbreviations used are: organization (o), biogenesis (b), establishment (e), maintenance (m) and assembly (a). Terms describing various types of metabolic and biosynthetic processes are omitted in the interests of space (native disorder is under-represented in these categories). The GO terms with names coloured in red annotate more than 200 proteins in yeast and have a p-value < 0.01 after the Bonferroni correction ($k = 45$).

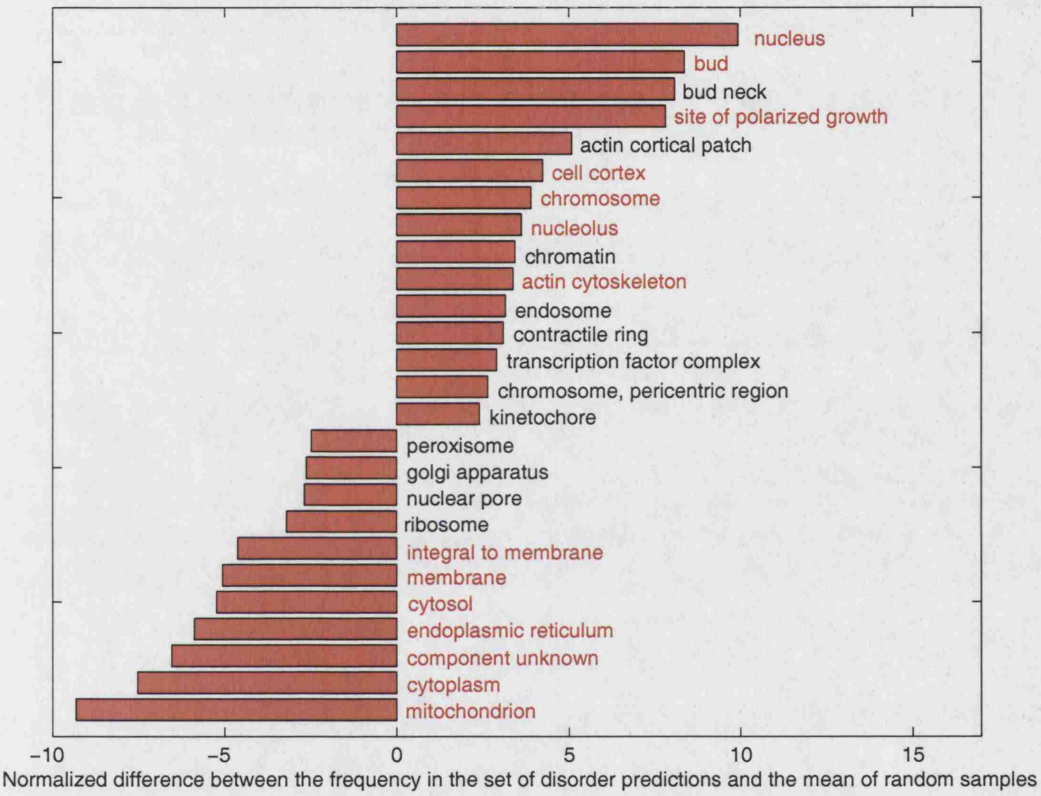

Figure 4.9: GO terms from the cellular component ontology that are significantly over- or under-represented in the set of disorder predictions. The GO terms with names coloured in red annotate more than 100 yeast proteins and have a p-value < 0.01 after the Bonferroni correction ($k = 32$).

4.6 Discussion

Many of the functions associated with disorder, mentioned in previous sections, are confirmed by Figure 4.7, which shows that the majority of putative disorder-containing proteins are involved in the molecular recognition of nucleic acids, nucleotides and other proteins. This is also reinforced by Figures 3.8, 3.9 and 4.10. Disorder is also associated with protein kinase activity, and since this is a regulatory process that requires simultaneous binding of a nucleotide and the protein phosphorylation site, it is consistent with the other functions that utilize disorder. The presence of disorder in kinases may also explain the small number of resolved crystal structures from this class of proteins. The low occurrence of disorder in functions such as biosynthesis and metabolism has also been indicated by previous work (Iakoucheva et al., 2002), and suggests that the rigid body model of molecular recognition applies fairly generally to the interactions between catalytic proteins and their substrates. The low frequency of disorder in catalytic proteins may also be one explanation for the preponderance of enzymes in the PDB with almost one-half of the entries belonging to this class of proteins (Hegyi and Gerstein, 1999).

Figure 4.9 indicates that disordered proteins tend to be located in cellular compartments that provide some protection from proteolysis such as the cell cortex and nucleus. The low levels of disorder in the mitochondrial proteins is consistent with the low frequencies of disorder in prokaryotes, as this organelle is believed to have originated as an infection of eukaryote cells by a purple-sulphur bacterium. The bacteria became fixed in eukaryotic cells, and the mitochondria have evolved endosymbiotically with the nuclear genome. Although the mitochondrion retains a small plasmid genome, many of the nuclear-encoded genes that are translocated to the mitochondria have evolved from those of the invading bacteria (Graur and Li, 2000; Schwartz and Dayhoff, 1978). Further support for the disparate disorder

frequencies between eukaryote and prokaryote proteomes is provided by Figures 4.9 and 4.8, which show that many of the cellular locales and functions that include proteins with putative disordered regions are unique to eukaryotes such as DNA packaging and cytoskeleton organization and biogenesis.

The presence of disorder in proteins that bind to the cytoskeleton explains its involvement in several processes that alter cell morphology in *Saccharomyces* such as polarized growth, budding and cytokinesis. A possible functional reason for the presence of disorder in cytoskeletal proteins is that the larger surface areas could increase the affinity with the surface of the cytoskeleton (Gunasekaran et al., 2003). Cytoskeletal proteins are also involved in signalling (Gundersen and Cook, 1999) and act as checkpoints at several stages of cell cycle. For example, inactivation of the tumour suppressor *p53*, which is known to contain unstructured domains (Lee et al., 1995), predisposes the cell to excess replication of the centrosomes and a failure to arrest cell cycle in the transition from G1 to S phase (Borel et al., 2002).

The experimental examples indicating that disordered proteins are involved in controlling gene expression, cell cycle, cell signalling and membrane fusion/transport, are also supported by Figure 4.8, which shows that these processes are also highly represented in the protein set predicted to contain long disordered segments. The predominant molecular function of long disordered segments appears to be binding of DNA to facilitate processes such as transcription, transposition, packaging, repair and replication. The presence of disorder in the RNA polymerase II complex indicates that the disordered binding domains are specific to the synthesis of mRNA for translation into protein (Graur and Li, 2000).

Several examples of natively disordered transcription activation domains have been discovered experimentally (Spolar and Record, 1994) and it is likely that a similar mechanism of protein and DNA recognition is used by the other processes.

Figure 4.8 also confirms that disorder is predicted in proteins that carry out membrane fusion and transport in endocytosis. The under-representation of disorder in the ribosome is explained by the lack of disorder in the structural constituents (Figure 4.7), although disorder does appear to be involved in the biogenesis and assembly of ribosomes via the processing of rRNA.

The results also suggest that disorder has a role in signal transduction via the small GTPases and cell surface receptors[7]. These pathways also facilitate responses to external stimuli, stress and the phases of cell cycle. Transformations from order to disorder may allow the cell to rapidly and semi-irreversibly reduce the concentration of signalling proteins in response to external or intracellular conditions. This is supported by experiments which show that proteolysis is a mechanism for inducing fast, semi-permanent changes during cell development (Nakayama et al., 2001). The uniform frequencies of disorder across eukaryote proteomes suggests that native disorder is also involved in higher functions beyond the basic control of cell cycle in unicellular yeast. A candidate in higher organisms is cell differentiation, which may involve disorder in controlling cell morphology *via* modifications of the cytoskeleton and controlling gene expression. Disorder may also be involved in the extrinsic signalling pathways that are present in multi-cellular eukaryotes such as apoptosis.

Native disorder has also been implicated in other cancer-associated proteins present in the Human genome (Iakoucheva et al., 2002). Figure 4.8 provides further detail on the causal mechanisms of cancer that may involve disorder such as gene silencing, epigenetic regulation of expression (Nephew and Huang, 2003) and DNA repair (Khanna and Jackson, 2001). The presence of disorder in Ty transposable element proteins also suggests that it is a feature of retroviruses since Ty elements are believed to have originated from a retroviral infection that has been fixed in the yeast genome. It is possible that the use of disorder for reversible binding of

[7]See supplementary information at http://bioinf.cs.ucl.ac.uk/disopred/suppInfo.html

DNA, and possibly transport through a small orifice (Daughdrill et al., 1998), is advantageous to retroviral infectivity in eukaryote cells.

In summary, native disorder is involved in some of the most important regulatory processes in eukaryotes; cell damage that renders some of these processes inactive is known to contribute to the development of cancer in humans. The presence of disorder in both tumour suppressors and oncogenes might therefore provide a novel target for future drug therapies. The abundant disorder in eukaryotes also has several implications for bioinformatics, and the prediction of protein structure in particular. The results of the previous chapter indicate that the folding of proteins into a compact structure is often incomplete in the absence of stabilizing proteins, DNA or ligands, and *ab initio* methods for predicting structure may therefore be improved by allowing for the flexibility that arises from disordered regions. Many obligatory domain-domain and protein-protein interactions are also likely to be mediated by co-operative folding of one or more of the proteins in the complex (Demarest et al., 2002), and this may also explain why the components are often difficult to crystallize in isolation. At the time of writing, some 150 proteins containing disorder have been described in the literature (Vucetic et al., 2005) and their importance in cell signalling, cancer and diseases caused by protein aggregation suggests that native disorder will be an active research area for the foreseeable future.

4.7 Future Work

There are two potential future directions for study of dynamically flexible protein structures that build on the research presented in the previous two chapters. The first is in improved prediction and characterization of disordered regions from experimental data. The second direction would involve gaining greater understanding of the evolution of disorder and its involvement in protein function on both molecular

and cellular levels. The following two sections describe some ideas for future work in greater detail.

4.7.1 Improving Prediction of Disorder

The previous chapter represents only a second attempt at disorder prediction and there are several modifications that could improve the prediction accuracy. Some of the residues in the set of structures used to train DISOPRED2 are recorded in the atomic co-ordinates and are therefore labelled as ordered, but have low occupancy and should be assigned to the disordered class. An example is shown in Figure 4.10. There is also a certain amount of subjectivity in whether crystallographers assign poorly resolved regions in the electron density map as missing co-ordinates or modelled residues with low occupancy or high B-factors.

The restriction to highly resolved protein structures, whilst rigorous in its definition of ordered protein structures, is likely to exclude more globally disordered structures, which will almost certainly be of poorer resolution. This results in predominantly short regions of disorder being recovered as shown in Figure 3.1. This is addressed to some extent by "hot loop" predictors but, as discussed previously, these may also be recovering very motile side-chains and not the flexible backbones representing dynamic disorder. A more systematic approach to defining disordered regions could make use of the electron density maps associated with lower resolution models. These could perhaps be used to discriminate between regions of the structure with multiple isoforms, low occupancy or poor model quality and those with the significant backbone flexibility that is characteristic of disorder.

Many of the structures in the training set for DISOPRED2 are also part of structural complexes and some regions of the component proteins may be disordered prior to the formation of the complex. For example, the predicted disordered region

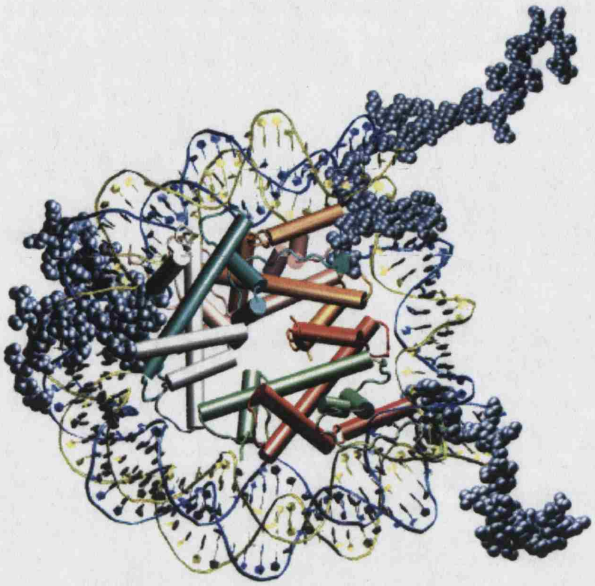

Figure 4.10: Histone proteins bound to DNA (1kx5). The disorder predictions are highlighted by the space-filling structure and the blue-grey colouring. The two predicted disordered regions that project outside of the complex were found to have zero occupancy in the PDB file.

on the left of Figure 4.10 appears to have formed a stable structure upon binding to DNA. A reduction in the number of training examples that are falsely classed as ordered could be obtained without restricting the size of the training set by labelling the regions in protein-protein and protein-DNA interfaces as unknown and using transductive learning algorithms, although these methods are computationally intense (Vapnik, 1998).

X-ray crystallography is only one of several experimental techniques that can be used to identify disordered proteins; circular dichroism, NMR spectroscopy and other indirect techniques from molecular biology such as differential proteolytic degradation are approaches that are not restricted to localized regions of disorder in predominantly ordered structures. The Dunker group has developed a database of

disordered proteins called *DisProt*, which is compiled manually from the literature (Vucetic et al., 2005). DisProt has some potential as a tool for enlarging the training set for disorder prediction, provided that the experimental definition of disordered regions is of sufficient quality. This would require further processing of the existing database as it currently includes examples of disorder that are characterized solely by CD spectroscopy, for example.

There are also several unusual protein sequences in DisProt that might impair the accuracy of any learning algorithm trained on the database. An example is the human form of the elastic *titin* protein, which is recorded as having a $\sim$ 2000 long region of disorder within its $\sim$ 7000 residue span. This giant protein is involved in muscle contraction, and has elasticity that is mediated by conformational change along its length. However, titin and its relatives are unique in terms of their size and function in higher eukaryotes. The assignment of disorder to the $\sim$ 2000-residue long PEVK segment has also been established using indirect experimental techniques (Granzier et al., 1996), and it is possible that some portions of this region are globular in their native state. Other examples of dubious disorder assignments occur in proteins that form part of the ribosomal complex.

Low sequence complexity is a characteristic of many disordered regions, and other methods for predicting disorder include a measure of entropy in the inputs to the classifier (Romero et al., 2001). It is not clear whether low sequence complexity is either;

1. a property of the amino acid sequence which *causes* structures to be disordered; or
2. an *effect* of the strong compositional biases that occur in disordered regions.

The first order entropy of the amino acid sequence S is calculated from the amino

acid composition for a window of residues

$$S = -\sum_{i=1}^{K}(n_i/N)\log(n_i/N) \tag{4.10}$$

where K is the size of the alphabet (in this case 20), n_i is number of times word i occurs in a window of length N. The window will therefore have a low sequence complexity, if the amino acid composition is biased toward a particular subset of residues.

The relatively high accuracy of the method based on the amino acid propensities for forming disordered structures (DISOcf) compared with the Dunker-Obradovic predictors suggests that this measure of sequence complexity does not improve the accuracy of disorder prediction. This fact, and the strong compositional biases shown in Figure 3.4 and independently by Romero et al. (2001), supports proposition (2). However, a recent review has suggested that the evolution of disordered regions is mediated by the repeated expansion of disordered sequence motifs (Tompa, 2003), which would suggest that the repetitive sequences are an evolutionary signature of disordered regions. This question could be investigated further by combining statistical analyses of amino acid composition with higher-order measures of sequence complexity. These measures, such as the linguistic complexity or k-order Markov complexity, could be used to investigate whether repetitive sequence motifs within a particular sequence are indicative of disorder (Orlov and Potapov, 2004; Lise and Jones, 2004).

4.7.2 Biological Applications

The estimates of disorder in complete archaean and eubacterial genomes provide very tentative evidence for disorder frequency being dependent on an organism's optimum

growth temperature. Simply extending the disorder frequency estimates to a larger number of prokaryote genomes could be used to investigate this dependence. A more refined investigation into the properties of disordered proteins might examine whether the presence of predicted disorder influences protein denaturation point or any other biophysical properties associated with the protein.

The presence of disordered regions in many DNA-binding sites also suggests that disorder predictions could be used to identify proteins that bind to DNA such as transcription factors. It may also be possible to develop DNA-disorder docking algorithms for finding the DNA sequence motifs that are complementary to a particular disordered region. However, the protein interaction interface samples a large number of structural configurations, is greatly influenced by the entropy contribution to the Gibbs free energy, and is likely to be difficult to model computationally. Disorder predictions could also be used to improve the prediction of other active sites such as those targeted for phosphorylation or ubiquitination, and to the determination of protein half-life. The most recent work by Iakoucheva et al. (2004) has shown that predictions of disorder can be used to improve the accuracy of sequence-based methods for recognizing protein phosphorylation sites. The results from this chapter also suggest that predictions of disorder could improve identification of domain boundaries.

Finally, the analysis using GO also suggests that the presence of long, disordered regions is linked to a variety of locations, functions and processes and may be useful for functional genomics. The following chapter describes the use of disorder predictions and other properties of amino acid sequences to predict the biological process annotations for yeast proteins.

Chapter 5

Ab initio Prediction of Protein Function in *S. cerevisiae*

The greatest challenge facing bioinformatics and computational biology lies in understanding how the individual components of biological systems act in concert to build and maintain the complexity of living organisms across many spatial and temporal scales. The vital first step of determining the DNA and the resulting protein sequences is at an advanced stage, and the structures of a large proportion of a typical proteome can now be determined accurately by homology with existing structures. The next step toward developing complete models of eukaryote cells is to infer the *functions* of the large number of gene products that have not been characterized by experiment.

The problem of assigning function to novel amino acid sequences has many similarities to structure prediction because of the causal link between structure and function. Both function assignment and structure prediction can be divided into different domains of difficulty based on the degree of similarity between the target sequence and the closest homologue from a database of proteins with known structure/function. The major difference between the two prediction problems is that whereas structure is a global property of the amino acid sequence, function can be dependent on the orientations of a small number of residues.

There is a strong evolutionary pressure on protein structures to be robust to random mutations (Xia and Levitt, 2004), but it is possible to alter the molecular function of a particular protein by directed mutation of several key residues. These include mutations of residues in post-translational modification or enzyme active sites. Targeted mutation of three or more residues can even alter the specificity of the interaction involving a relatively large protein-protein interface (Kortemme et al., 2004). As a consequence, pairs of sequences with PSI-BLAST expectation values below 10^{-3} usually adopt similar structures (Rost, 2001b) but even matches below 10^{-50} do not always share identical biochemical functions (Rost, 2002).

The other important distinction between structure and function prediction is that most aspects of protein structure have a relatively straightforward physical definition (Native disorder is a notable exception) but function can be defined by numerous properties of the protein such as its biochemical activity, its interactions with other macromolecules within the cell or its overall effect on the organism. Successful computational techniques for inferring each aspect of protein function are therefore likely to be based on different sources of information and biological principles.

The molecular function of a protein depends on the chemical properties and orientations of a relatively small number of residues in the protein structure but biological process is dependent on context, including factors external to the amino acid sequence such as gene expression, cellular location and protein degradation, activation and inhibition. Indeed, the key genomic discoveries of the past year have indicated that mutations in regulatory regions contribute much more to genetic and morphological diversity than changes in the amino acid sequence. Examples include the observation that many apparent non-coding regions have extreme (ultra) conservation across widely divergent chordate species (Bejerano et al., 2004) and that repeat expansions in cis-regulatory regions are responsible for much of the morphological variation between breeds of the domestic dog *Canis familiaris* (Fondon and Garner, 2004).

This chapter investigates the utility of phylogenetic profiles (Marcotte et al., 2000) and simple sequence-derived features such as amino acid composition and secondary structure/disorder predictions (Jensen et al., 2003) for inferring different aspects of protein function. These different aspects of protein function are defined using the three separate ontologies that comprise the Gene Ontology (GO). This chapter therefore describe the use of support vector machines (SVMs) to investigate which of the different aspects of protein function can be predicted most accurately

using phylogenetic profiles and simple composition-based features.

5.1 Inferring Protein Function: A Review

The majority of the previous studies which used supervised learning algorithms for predicting protein function did not make a distinction between proteins that have similar sequences to a functionally annotated homologue and those that do not (Marcotte et al., 1999b; Pavlidis et al., 2002; Lanckriet et al., 2004). This is unlikely to be an optimal approach, as experience of the related structure prediction problem has shown that different techniques are required in the two cases.

The development of more accurate methods for inferring function *via* homology is likely to involve incorporating prior knowledge of the evolution of function in specific protein families, the conservation of active site residues within a particular family, and how the biochemical activities of single domains are combined to create novel functions in multi-domain proteins (Todd et al., 2001; Abascal and Valencia, 2003; Pazos and Sternberg, 2004). However, this type of approach does not build on the work of the previous three chapters, which have concentrated on using supervised learning to predict *ab initio* protein structure. The objective of this chapter is therefore to predict the functions of proteins that do not have strong similarity with proteins that have been characterized by experiment. It is therefore desirable to structure the cross-validation experiment so that the performance is as close as possible to what would be expected in practice (i.e. for proteins without a functionally-assigned homologue).

However, the Gene Ontology is a fairly recent development and there has been relatively little work on assigning GO terms using sequence similarity. For example, it is not known at what level of sequence similarity a biological process (BP) annota-

tion can be transferred accurately to a similar orthologous or paralogous sequence. The most comprehensive studies of the relationship between sequence similarity and function conservation have concentrated on investigating the number of shared Enzyme Commission (EC) numbers[1] between sequences at varying thresholds of sequence similarity (Wilson et al., 2000; Todd et al., 2001; Rost, 2002).

These three studies have produced varying estimates of the accuracy of annotation transfer, with Rost (2002) arguing that the results of Todd et al. (2001) and particularly Wilson et al. (2000) are optimistic (i.e. significantly higher than would be expected for annotating a typical proteome) because of the biased nature of the sampling used to generate experimental annotations for SWISSPROT or to obtain protein structures for deposition in the PDB. The investigation by Rost (2002) makes an attempt to correct for this bias, and is based on a larger number of sequence comparisons (10^7 compared with $10^3 - 10^4$). The results indicated that less than 30% of sequence pairs sharing $> 50\%$ sequence identity were described by identical EC numbers, suggesting that protein structure can be determined 'trivially' at much lower levels of sequence similarity than biochemical function (Rost, 2002).

The studies based on EC numbers are limited to a subset of all possible biochemical functions, and to proteins with a single EC annotation (usually proteins consisting of a single domain). Although EC numbers have the advantage that they provide a simple description of the protein's catalytic activity, they do not account for other functional characteristics of the protein, such as the pathway or broader BP in which it participates. The following section describes some of the limitations of sequence similarity and structure predictions for inferring these and other aspects of protein function.

[1]The EC numbers are a hierarchical scheme for describing the reactions catalyzed by natural enzymes.

5.1.1 Limitations of Sequence Similarity and Structure Prediction for Establishing Function

The standard means of annotating the function of a newly-uncovered protein is to perform a sequence search against a database of proteins with known functions, and to assign the novel sequence the same functional label as the most similar sequence in the knowledge base. This approach can lead to the occurrence of a large number of annotation errors (Brenner, 1999), and as a result the curated sequence databases such as UniProt and the SGD do not provide annotations that are assigned solely by homology (Apweiler et al., 2004; Dwight et al., 2002).

The mapping from structure to function is also complicated, and it is possible for proteins that share almost identical structures to perform different functions. Another possibility is that identical functions can be carried out by proteins that do not have homologous structures. This phenomenon is known as convergent evolution, as different structural solutions have evolved independently to perform the same function with the most famous examples being the serine proteases subtilisin and trypsin. However, these two examples are fairly exceptional, with most structurally similar proteins performing related biochemical functions.

There are also several more important factors that limit the scope of using homology for inferring various aspects of protein function. The first is the case where a protein does not share any sequence similarity with existing structures. The other related but more subtle problem occurs when only a partial alignment is recovered between a novel sequence and a sequence with known function. This could, for example, indicate that a protein consisting of two domains contains a DNA binding domain but provide little information regarding the activity of the other domain.

The other more difficult approach to inferring function is to proceed *via* the

intermediate step of predicting the protein's structure. Fold recognition methods (McGuffin and Jones, 2003) have increased sensitivity over sequence searching by matching a particular sequence to a structural template. Although there is often a large degree of functional diversity at this level of structural similarity, knowledge of the fold reduces the number of potential functions of the protein. Even the structural class of the protein contains a signal for discriminating different protein functions. For example, the majority of enzymes are α/β proteins (Martin et al., 1998). It is also possible to use the three-dimensional models from structural genomics targets (Laskowski et al., 2003; Pazos and Sternberg, 2004), homology modelling, fold recognition (Sodhi et al., 2004a) or *ab initio* structure prediction to infer the biochemical activity of the protein from its structure. However at the present time, the poorer quality *ab initio* and fold recognition models may not to be sufficiently precise to allow accurate predictions of molecular function.

The other deficiencies of structure-based function predictions are that many types of protein-protein (Wodak and Mendez, 2004) and protein-DNA interactions cannot be predicted accurately or efficiently, and that a large number of protein structures cannot currently be determined using existing experimental or computational techniques. This includes many membrane proteins, which account for around one-third of a eukaryote proteome. Membrane proteins have different folding patterns as a result of the hydrophobic environment of biological membranes, and have structures that are difficult to solve using X-ray crystallography. However, the clinical importance of membrane proteins such as cell surface receptors and ion channels as drug targets, has led to the establishment of structural genomics consortia that are dedicated to improving the technology for solving the structures of membrane proteins (Kyogoku et al., 2003).

There is also a growing acceptance that the structure-function paradigm, which has prevailed in biochemistry for the previous couple of decades, may need to be

modified to account for the one-fifth of a typical eukaryote proteome that is estimated to contain long regions of disorder (Wright and Dyson, 1999; Ward et al., 2004b). The proportion of protein structures that can be determined readily using conventional techniques is reduced still further by proteins in large complexes, and in particular, the cytoskeleton. The limitations of these techniques and the slow and labour-intensive nature of experimental biology, has stimulated the development of function prediction methods that do not rely directly on sequence similarity or knowledge of a protein's structure. The methods that are relevant to the latter half of this chapter are reviewed in the following section.

5.1.2 The Phylogenetic Profile

One of the approaches to annotating protein function which makes use of the abundant sequence data is the phylogenetic profile (Pellegrini et al., 1999). A protein's phylogenetic profile is a vector of scores which represent the degree of sequence similarity between the protein and the closest homologue recovered from the proteome of another organism. In the simplest representations, this is a binary vector with unity representing the presence of another sequence above a particular similarity threshold, and zero the absence of such a sequence. These bit vectors $\mathbf{x} = (x_1, \ldots, x_m)'$ have a dimensionality equal to the number of genomes included in the comparison, m.

The matrices $\mathbf{X} = (\mathbf{x}_1, \ldots \mathbf{x}_n)$ of gene content for a set of n protein sequences can also be used to construct phylogenetic trees for organisms (Huson and Steel (2004); Gu et al. (2005); Figure D.1). However, this section concentrates on using the gene loss/conservation vectors of yeast proteins to predict their GO annotations. This use of phylogenetic profiles can be viewed as a global approach to comparative genomics, which has traditionally been used to identify genes that differ between

closely related species. An example is the comparative genomic approaches to the identification of proteins involved in bacterial pathogenicity. In these studies, the proteome of a pathogenic bacterium is compared to a non-pathogenic relative on the assumption that potential mediators of infectivity or toxicity are likely to be unique to the pathogen or have significant differences from the most similar orthologous sequences in the non-pathogen (Glaser et al., 2001).

The earliest method for generating profiles performed global alignments between each protein in *E.coli* and the proteins from 16 other completed genomes (Pellegrini et al., 1999). The statistical significance (p-value) of each match was then calculated by performing random shuffles of the sequence and assuming a Gaussian distribution of optimal alignment scores. Matches with an expectation value, E[2], below 1 were used to define the presence of a homologous sequence. Pellegrini et al. (1999) also showed that 'neighbouring' profiles, defined as having a Hamming distance of less than 3, tended to share EcoCyc[3] annotations and were described by similar SWISSPROT keywords.

This was followed by a second method (Marcotte et al., 1999b), which measured the correlation between the binary phylogenetic profiles of all yeast genes to establish functional links between proteins with highly-correlated profiles. Finally, Marcotte et al. (2000) calculated the profile score for each genome using $-1/\log(E)$, where E is the expectation value of the closest homologue recovered from a BLAST search. These scores were set to one for $E > 10^{-6}$. Marcotte et al. (2000) also used the profile scores and Fisher's linear discriminants to predict sub-cellular location and to infer whether a set of mitochondrial proteins originated in prokaryotes, eukaryotes or were orphans (unique to a single organism).

[2] $E = nmp$ where p is the p-value, and m and n are the number of proteins in *E.coli* and the 16 comparison genomes.

[3] Encyclopedia of Escherichia coli Genes and Metabolism.

Since these initial studies, various continuous-valued profiles constructed from modified BLAST E-values have been shown to provide greater information than the binary vectors (Enault et al., 2003). However the finding by Enault et al. (2003) that correlation provides higher accuracy than inner products is surprising given the relationship between the two operators (Appendix C). It also seems likely that simple distance measures, such as correlation, will lead to the loss of a significant amount of information that is contained in the profile vector. For example, it would be expected that a protein whose main role was in modifying actin would only have homologues in other eukaryote organisms. Methods that make use of higher level properties of the phylogenetic profiles such as presence or absence in a particular class of organism may improve upon the simple functional linkage approach.

To date, two other groups have used SVMs for predicting protein function from phylogenetic profiles. Pavlidis et al. (2002) developed a method for integrating phylogenetic profile and microarray data into an overall prediction of function of yeast proteins as defined using the MYGD[4] classifications. Pavlidis et al. (2002) generated the profiles by taking the negative logarithm of the closest homologue recovered from the BLAST search.

Although the authors recognized that correlated profiles could arise trivially as a result of sequence similarity, they did not structure the cross-validation to ensure that similar sequences did not appear in test and training sets simultaneously. There are two reasons for this approach being undesirable. Firstly, the performance of the prediction method is not evaluated under conditions that resemble how it is likely to be used in practice (see Section 5.2.1 for further discussion). Secondly, phylogenetic profiles are based on the assumption that certain pathways are inherited as intact functional modules throughout the process of evolution. This biological hypothesis is not tested for classes that only contain homologous sequences, as these proteins

[4]Munich Information Center for Protein Sequences Yeast Genome Database.

would recover similar profiles from any random partitioning of a large sequence database. This resulted in several small homogeneous classes such as the amino acid transporters and chaperones being recovered with highest accuracy.

Another method used a very similar experimental set-up, except binary profiles were generated by thresholding the BLAST expectation values at an E-value of 1 (Vert, 2002). These were classified using a linear kernel and a "tree kernel", which was constructed as a generative model for gene loss/conservation on the phylogenetic tree of 24 organisms in the comparison. Although the tree kernel is a novel and innovative use of generative models in functional genomics, the choice of threshold greatly impairs the accuracy of the method. This occurs because a divergent sequences (e.g. with an E-value of 0.9) and highly similar sequences are both recorded identically as sequence-similar orthologues in the binary profile (see Section 5.2). The study also failed to structure the cross-validation so that homologous proteins did not occur simultaneously in both test and training sets.

5.1.3 Inferring Protein Function using Sequence Compositional Features

The simplicity of representing protein sequences using their averaged amino acid composition has resulted in these measures being used in almost every area of protein sequence/structure analysis, from the prediction of native disorder (Romero et al., 1997) to sub-cellular location (Park and Kanehisa, 2003). Amino acid composition counts the relative frequencies of words in the protein sequence of length 1. Words of greater length such as PROSITE motifs (Hulo et al., 2004), which are associated with a specific type biochemical activity, could also be used to infer protein function. In fact, the protein domain complement could be viewed as a grammar for specifying the function of a protein. Although previous results using this approach do not appear

to be particularly promising (Cai and Doig, 2004), a domain projection method has been shown to be accurate in predicting the subcellular location of a subset of all proteins with very high accuracy (Mott et al., 2002).

The publication most relevant to this work (Jensen et al., 2003) used amino acid composition and simple sequence-derived features such as net charge, hydrophobicity, isoelectric point and predictions of signal peptides at the N-terminus to predict several GO biological process classes for proteins from the Human genome. This study used selection methods to identify which features were useful for discriminating particular function classes, and trained radial basis function networks to solve the classification problem (the authors noted that SVMs produced lower accuracies than the RBF networks). The study by Jensen et al. (2003) also structured the cross-validation experiment so that no homologous proteins occurred in both test and training sets. The following section describes a similar system for assessing the effectiveness of phylogenetic profiles and compositional features for inferring protein function.

5.2 System and Methods

One complication of using the Gene Ontology to investigate the function of amino acid sequences is that there is significant redundancy in the terms used to describe each gene product. This is demonstrated by the total of 3380 valid GO terms attached to sequences within the yeast proteome. This redundancy was encountered in the investigation of disorder with the related terms of 'kinase activity', 'protein kinase activity' and 'protein serine/threonine kinase activity' found to be similarly enriched in long disordered segments because they describe almost identical sets of yeast proteins (see Chapter 4). This was overcome in an *ad hoc* manner by removing very similar terms from Tables 4.7-4.9. However, this approach is made

unfeasible for automatic prediction by the very large number of terms in the entire Gene Ontology. An alternative is to use the 'slims' version of the ontology (Gene Ontology Consortium, 2001), which provide large, general categories of protein function. However, this chapter is concerned with identifying smaller functional classes that show similar patterns of inheritance across complete genomes. The following section therefore describes a protocol for removing redundant GO terms and ensuring that homologous proteins do not occur simultaneously in test and training sets.

5.2.1 Partitioning the Data Set and Selecting Representative Classes from the Three GO Ontologies

Figure 5.1 shows a schematic representation of the protocol for selecting terms from the molecular function (MF), biological process (BP) and cellular component (CC) ontologies for *ab initio* prediction of protein function.

1. The yeast sequences and their gene associations were downloaded from the February 2004 release of the Saccharomyces Genome Database. These are divided into the 5881 ORFs that have been verified experimentally or by orthology with the genomes of closely related yeast species and 822 "dubious" ORFs. All dubious ORFs were excluded as, by definition, these sequences do not have high quality experimental annotations. Table 5.1 shows the number of protein sequences from the set of 5881 verified ORFs that are assigned at least one MF, BP or CC annotation. There were a total of 3380 GO terms used to annotate proteins in yeast. This number includes the primary annotations found in the gene association file and all ancestral nodes in the GO directed acyclic graph. The ancestral terms provide valid but more general descriptions of the protein's function.

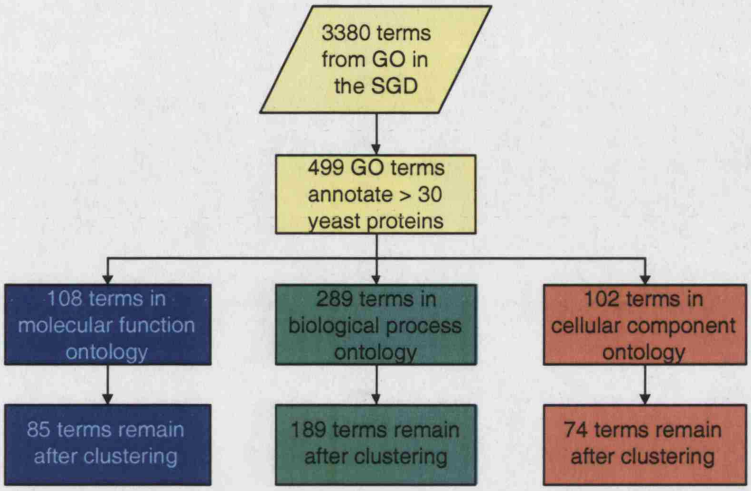

Figure 5.1: Flow diagram of the protocol for selecting GO terms from the molecular function, biological process and cellular component ontologies for *ab initio* prediction of protein function. In the February 2004 release of the SGD, there were a total of 3380 GO terms used to describe the functions of proteins in *Saccharomyces cerevisiae* (excluding the terms Gene Ontology, molecular function ontology, biological process ontology and cellular component ontology). The terms annotating fewer than 30 yeast proteins were removed, and the remaining terms were divided into the three separate ontologies. Finally, these terms were clustered using the Tanimoto metric on the sets of yeast proteins annotated by each term. A complete description of each step is described in the text.

	Annotated	Unknown
Molecular function	3539	2342
Biological process	3094	2787
Cellular component	4884	997

Table 5.1: Number of verified ORFs in *Saccharomyces cerevisiae* with at least one annotation from the molecular function, biological process and cellular component ontologies

2. Of the 3380 GO terms in the SGD, 499 annotate more than 30 yeast proteins. This threshold was chosen to include as large a number of functional descriptions as possible without resulting in class membership ratios that greatly exceed 100:1. In later sections, binary classifiers are trained to discriminate between sequences with a particular GO annotation and all other yeast proteins that are assigned an annotation from the same ontology (the annotated sequences in Table 5.1). Of the 499 terms, 108 terms are from the MF ontology, 289 from the BP ontology and 102 from the CC ontology.

3. There is significant redundancy in the set of terms from each ontology. For example, the terms "sensory perception", "perception of external stimulus", and "perception of chemical substance" describe an identical set of yeast proteins, although this will not be the case for higher organisms. More generally, GO terms which describe very similar sets of proteins will be recovered with similar accuracy (previous chapter). This complicates assessing the performance of each classifier and interpreting the results. The redundancy was therefore reduced by clustering the GO terms using the Tanimoto distance between the sets of yeast proteins associated with each term. The Tanimoto distance between two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ is defined as

$$D_{\text{Tanimoto}}(\mathcal{S}_1, \mathcal{S}_2) = \frac{n_1 + n_2 - 2n_{12}}{n_1 + n_2 - n_{12}} \tag{5.1}$$

where $n_1 = |\mathcal{S}_1|$, $n_2 = |\mathcal{S}_2|$ and $n_{12} = |\mathcal{S}_1 \cap \mathcal{S}_2|$. The Tanimoto distance is particularly useful for taxonomic problems, such as this, because it satisfies all the properties of a metric (non-negativity, reflexivity, symmetry and the triangle inequality) and can therefore be used in conjunction with standard clustering algorithms. Figure 5.2 shows the relationship between the minimum Tanimoto distance threshold and the number of BP clusters recovered from several hierarchical clustering algorithms. The absence of any structure in

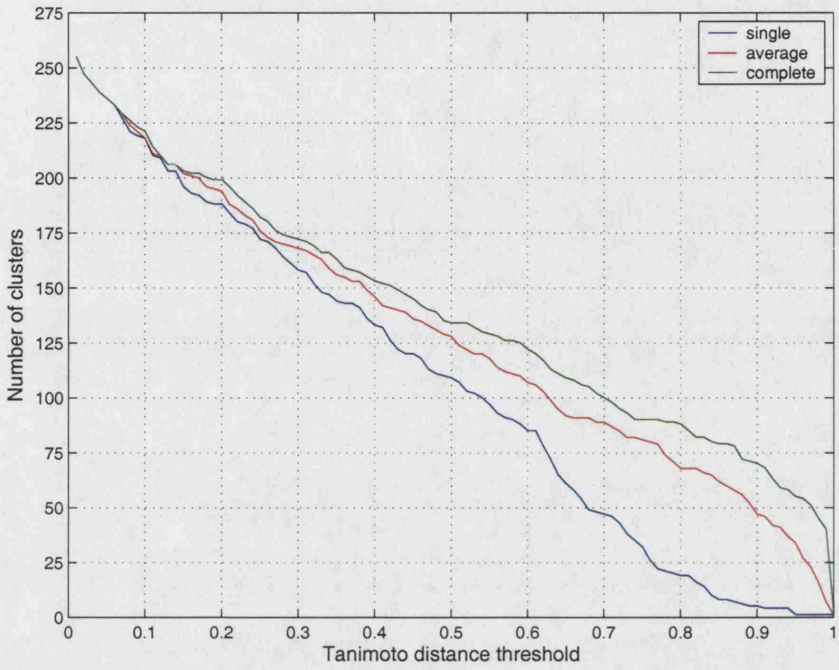

Figure 5.2: Number of clusters containing GO terms as a function of Tanimoto distance threshold for several clustering algorithms. The results are from hierarchical clustering algorithms which, at each iteration, combine the two clusters that are separated by the shortest distance into a single cluster. The single linkage criterion is the shortest distance between the members of a cluster and another candidate cluster, the average is the mean distance, and 'complete' is the largest distance between members of the two clusters.

Figure 5.2 led us to select a threshold Tanimoto distance of 0.2, as this removes much of the redundancy in the set of GO classes without greatly compromising the descriptive power of the remaining terms (similar plots are obtained for the MF and CC ontologies). For example, two sets of identical size $|\mathcal{S}|$ have a Tanimoto distance of 0.2 if their intersection is approximately $0.89|\mathcal{S}|$.

The GO term that annotated the largest number of yeast proteins was selected from each cluster generated using the single linkage distance. In cases where several GO terms were used to describe identical sets of yeast proteins, the

term which provided the most appropriate subjective description was selected. In the previous example, the term "perception of a chemical stimulus" was used as this provides the best description of the process of sensory perception in a unicellular organism such as yeast.

4. In the final step, the amino acid sequences in yeast were clustered by sequence similarity. This was achieved by performing a three-iteration PSI-BLAST search for each yeast protein against a large, non-redundant database that also contained all sequences from the SGD. Two yeast sequences were considered similar if they shared more than 25% sequence identity. This threshold is similar to that used for benchmarking secondary structure prediction and for the assessment of fold recognition methods (McGuffin and Jones, 2003). The threshold is also justified on the grounds that enzyme function is not well conserved below sequence identity thresholds of 30% (Todd et al., 2001) and that it is consistent with other work in the field (Jensen et al., 2003).

   The graph of similarity relationships (edges) between proteins (nodes) was then used to obtain homologous clusters by placing all connected nodes in a single cluster. The pair-wise similarity scores were used to obtain equivalence classes using tree-generating code from *Numerical Recipes* (Press et al., 1988). This clustering algorithm obtains the equivalence classes efficiently by generating a tree structure for each equivalence class. A pair of proteins are placed in the same equivalence class by making one protein in the pair a parent of the other. Further objects are added to each equivalence class by adding these members to the tree structure.

The following section describes the system for predicting protein function using phylogenetic profiles, and investigates the optimal representation of the profile vector for the prediction of BP classes.

5.2.2 System for Investigating the Optimal Representation of the Phylogenetic Profile Vector

A total of 93 eukaryote and prokaryote proteomes were used to generate the phylogenetic profile. These were obtained from the NCBI ftp server and placed in a single flat file. This large sequence database was filtered using the sequence masking program *pfilt* to remove putative coiled-coil, transmembrane and low complexity regions (Jones and Swindells, 2002). The sequences from the SGD were then queried against the database with the BLAST jobs distributed across a Linux Beowulf cluster of Intel Pentium and AMD Athlon processors and two associated SunFire 880 servers running Solaris. The profile vector for each yeast sequence was constructed using the closest match (defined using E-value) from each of the 93 genomes, shown in Appendix D.

The literature review in Section 5.1.2 described several representations of the phylogenetic profile vector that have been used to predict protein function (Marcotte et al., 2000; Vert, 2002; Pavlidis et al., 2002). These profile representations include normalized bit-scores [5, 6] from a single iteration BLAST search and binary profiles

[5] In all cases where the profile vectors are described as 'normalized', the vectors have been linearly scaled to have unit Euclidean lengths.

[6] Although Pavlidis et al. (2002) used the negative logarithm of the BLAST expectation-value, the bit score is more precise as it is proportional to the negative logarithm of E but is not subject to the loss of numerical precision associated with calculating and then inverting an exponent

$$E = \frac{mn}{2^{S'}} \tag{5.2}$$

$$\log(E) = \log mn - S' \log(2) \tag{5.3}$$

where S' is the bit-score, and m and n are the effective lengths (in residues) of the target sequence and the search database. For a particular query protein, the BLAST bit-score is therefore linearly related to the logarithm of the BLAST E-value.

which are generated by thresholding the BLAST expectation values at 10^{-6} and 1.

The success of these different representations and the implications for conservation of intact functional modules between genomes was investigated by training linear SVMs to recognize each of the 189 BP classes described in Section 5.2.1. The BP ontology was chosen to investigate the optimal profile representation, as it is this aspect of protein function which is believed to be conserved between proteins that have a similar pattern of inheritance across complete genomes. The SVMs were trained, using the SVMlight package (Joachims, 1999), to discriminate between sequences within an individual BP class and all other yeast proteins with the exception of those sequences that are annotated with "process unknown". The frequency imbalance of many of these classification problems was dealt with by setting the asymmetric cost parameter $j = C_+/C_- = N/2N_p$ where N and N_p are the number of examples in the training set and the number of positive examples respectively (see Chapter 3).

The performance of each SVM was evaluated by calculating the area under the ROC curve (Wilcoxon statistic or ROC-score) from the pooled results of 4-fold cross-validation experiments. The Wilcoxon statistic is used as this is the only measure of accuracy that is not dependent on the class frequencies. Each cross-validation experiment was repeated with 10 random partitions of the training data to obtain estimates of the mean and standard deviation of each ROC score, and the cross-validation was structured so that profiles from homologous sequence clusters did not appear simultaneously in both training and test sets (see Section 5.2.1).

The success of the three methods for encoding the phylogenetic profile of a protein sequence is presented in the following section, which also investigates whether using PSI-BLAST to detect more remote homologues in the query proteomes improves the prediction of biological process.

Encoding scheme	Encoding scheme	C_f	μ_d	p
BLAST (bit-score)	PSI-BLAST (bit-score)	0.85	-0.015	4.9×10^{-6}
BLAST (bit-score)	BLAST (10^{-6})	0.68	0.027	1.5×10^{-8}
BLAST (bit-score)	BLAST (1)	0.78	0.025	2.0×10^{-9}
BLAST (10^{-6})	BLAST (1)	0.75	-0.002	0.61

Table 5.2: Comparing several schemes for representing phylogenetic profiles for the biological process class prediction problem. C_f is the correlation between the ROC scores for each BP class prediction problem using the two alternative encoding schemes. μ_d is the mean difference between the encoding scheme in the first and second columns $\mu_d = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)$ and p is the probability of observing the variables x_i and y_i under the null model ($\mu_d = 0$), calculated using a paired t-test.

Results

Figures 5.3-5.5 show comparisons of the ROC scores produced by linear SVMs on the BP class prediction problem. Each point on these plots represents the area under the ROC curve for predicting a single BP class using the encoding scheme plotted on the x- and y-axes. Figure 5.3 shows a comparison between the ROC scores for predicting BP classes using normalized bit-scores from BLAST and three-iteration PSI-BLAST searches. Figures 5.4 and 5.5 compare normalized bit-scores, and binary scores from thresholding the E-values from single-iteration BLAST searches.

The blue line, plotted in each of the Figures, shows the trend line that would be expected if the two methods in the comparison provided equal accuracy ($x = y$). Qualitatively, a large proportion of points below the blue line indicates that the encoding scheme plotted on the x-axis has highest accuracy. The red line indicates the linear least-squares fit to the 189 data points plotted in each figure. Table 5.2 presents other results from the three comparisons.

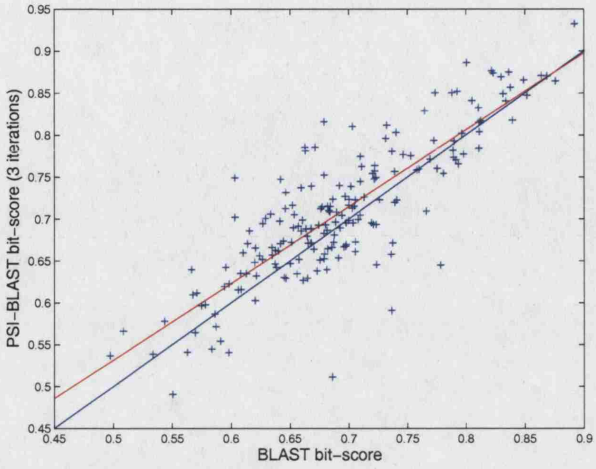

Figure 5.3: Comparison of ROC scores for predicting biological process classes with linear SVMs trained on two different representations of the phylogenetic profile vector. The profiles are encoded using normalized bit-scores, which are calculated from BLAST and three-iteration PSI-BLAST searches. The red line shows the linear least-squares fit and the blue line $y = x$.

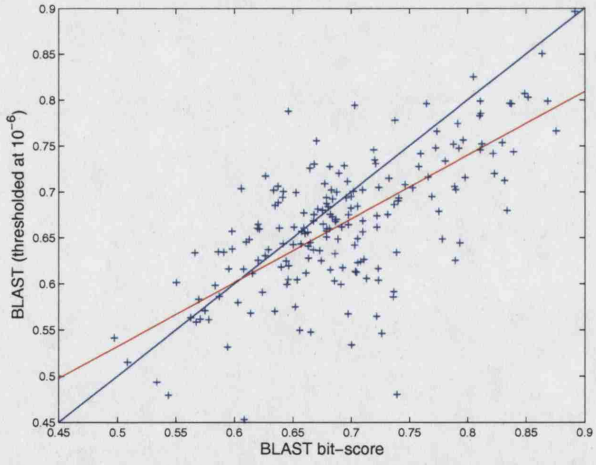

Figure 5.4: Comparison of ROC scores for predicting biological process classes with linear SVMs trained on normalized bit-scores and binary scores generated by thresholding E-values at 10^{-6} from the same single-iteration BLAST search.

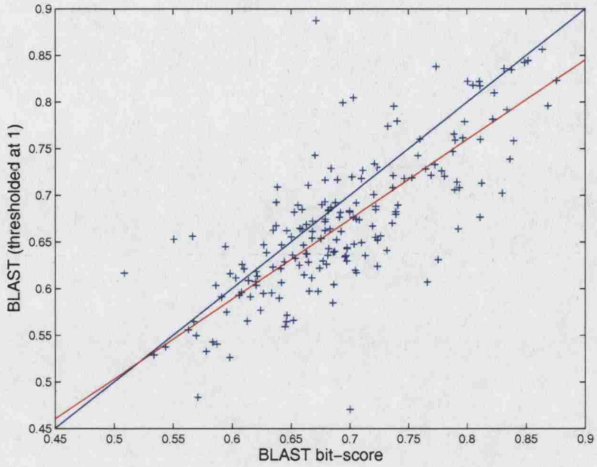

Figure 5.5: Comparison of ROC scores for predicting biological process classes with linear SVMs trained on normalized bit-scores and binary scores generated by thresholding E-values at 1 from the same single-iteration BLAST search.

Figure 5.3 indicates that phylogenetic profiles, constructed using bit-scores from PSI-BLAST searches can be used to predict biological process classes with slightly higher accuracy than profiles generated using BLAST. Figures 5.4 and 5.5 show that thresholding the BLAST E-value leads to a reduction in prediction accuracy.

It would be expected that the BLAST and PSI-BLAST-generated profiles would be highly correlated, since the increased sensitivity of profile-based sequence search methods arises from the use of information from sequences with intermediate similarity (Altschul et al., 1997; Park et al., 1998). As each feature in the phylogenetic profile represents only the most similar sequence in the comparison genome, the presence of more distant homologues within that genome will not be recorded in the profile. Therefore, the increase in the overall accuracy of classifiers trained on PSI-BLAST-derived phylogenetic profiles is likely to arise from the inclusion of more remote homologues from the other genomes in the comparison. However, the relatively large variation in the accuracies between predictions based on BLAST and

PSI-BLAST searches (Table 5.2, Figure 5.3), suggests that sequences detectable by PSI-BLAST but not BLAST, do not necessarily represent functionally conserved orthologues in the comparison genome, and perhaps reflects function being conserved at different levels of sequence similarity within different process classes. The other comparisons indicate that a measure of the level of similarity between a protein and the closest orthologous sequence improves the accuracy of BP prediction. Slightly surprisingly, the results do not depend greatly on the threshold used to generate the binary phylogenetic profiles. This may be caused by there being relatively few homologues recovered at E-values between 10^{-6} and 1.

The following section investigates the use of disorder predictions and other simple sequence-derived features for inferring protein function.

5.2.3 Effect of Native Disorder on the Accuracy of Predicting Classes from the Three GO Ontologies

The results from Chapter 4 indicated that disorder predictions were associated with numerous molecular functions, biological processes and cellular locales. This section therefore investigates whether predictions of disorder improve the accuracy of function prediction over those based on other simple compositional features of the amino acid sequence. This is tested by training linear SVMs to classify terms from the three GO ontologies using the amino acid composition, which is calculated by counting the frequency of each of the twenty amino acids and dividing by the length of the sequence, the secondary structure composition predicted using PSIPRED (Jones, 1999), and the disorder composition predicted using DISOPRED2 (Ward et al., 2004b). The PSIPRED and DISOPRED2 prediction methods were both run using standard parameters with the DISOPRED2 false positive rate threshold set to 5%. There were therefore a total of 24 inputs encoding the composition of the

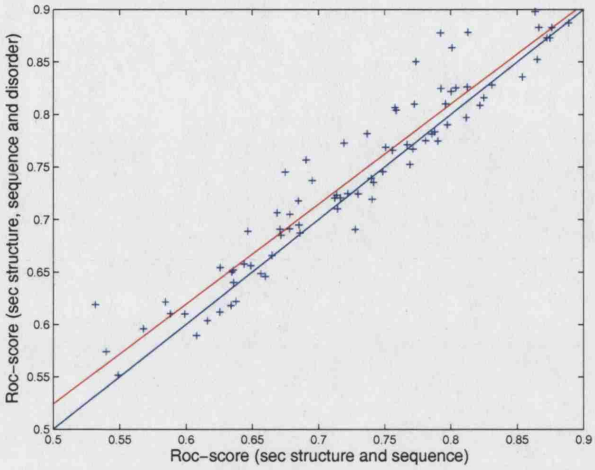

Figure 5.6: Comparison of ROC scores for predicting molecular function classes with linear SVMs trained on all compositional features and SVMs trained with the predicted disorder composition excluded.

20 amino acids, in addition to features representing the helix, sheet, coil and disorder composition. The protocol is identical to that used to predict GO classes using phylogenetic profiles, except with the profile vectors replaced with the 24 composition-based features.

Figures 5.6-5.8 show comparisons of the ROC-scores for predicting MF, BP and CC classes using linear SVMs trained on all of the compositional features, including predictions of the disorder composition, and linear SVMs trained on only sequence composition and predicted secondary structure composition. As before, the blue line is $y = x$ and the red line indicates the linear least-squares fit to the data points plotted in each figure. The Table 5.3 presents other results from the three comparisons.

The results in Figures 5.6-5.8 and Table 5.3 indicate that predictions of the disorder composition from DISOPRED2 provide a signal for identifying the three aspects of protein function defined by the Gene Ontology. The greatest improvements occur

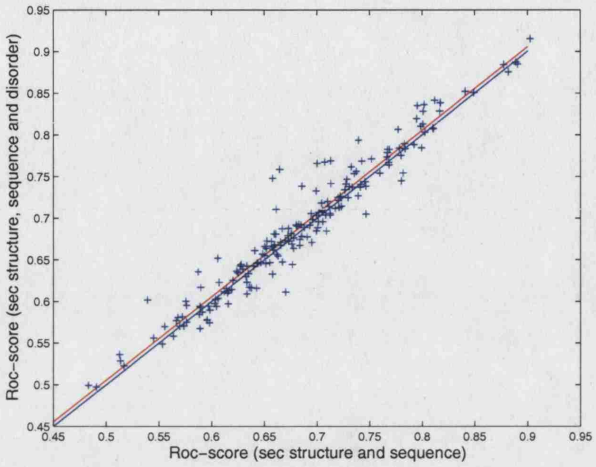

Figure 5.7: Comparison of ROC scores for predicting biological process classes with linear SVMs trained on all compositional features and SVMs trained with the predicted disorder composition excluded.

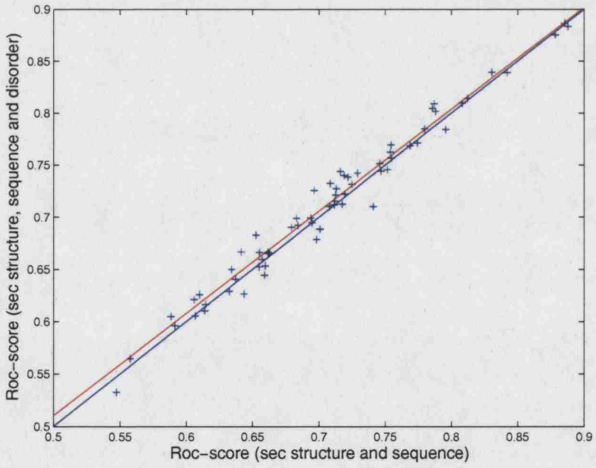

Figure 5.8: Comparison of ROC scores for predicting cellular component classes with linear SVMs trained on all compositional features and SVMs trained with the predicted disorder composition excluded.

Ontology	C_f	μ_d	p
Molecular Function	0.967	12.96×10^{-3}	1.72×10^{-5}
Biological Process	0.970	5.69×10^{-3}	1.80×10^{-4}
Cellular Component	0.990	5.73×10^{-3}	6.70×10^{-4}

Table 5.3: Comparison of linear SVMs trained on all of the sequence compositional features and those features with predicted disorder composition excluded. μ_d is the mean difference in the ROC scores based on all inputs, x_i, compared with the same inputs with disorder excluded, y_i. $\mu_d = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)$ and p is the probability of observing the variables under the null model ($\mu_d = 0$), calculated using a paired t-test. C_f is the correlation between the ROC scores for predicting classes from the three ontologies using the two alternative sets of inputs.

in predictions for the molecular function ontology, which is consistent with disorder's primary role being in molecular recognition. This provides further tentative evidence for structure-based molecular function prediction (Laskowski et al., 2003) being improved by incorporating some degree of back-bone flexibility in the definition of the active site region.

The following section describes the protocol for assessing the utility of the compositional features described in this section and phylogenetic profiles for providing guidance for further experimental studies of protein function.

5.2.4 Predicting Classes from the Three GO Ontologies

Although the area under the ROC curve provides a measure with a solid statistical foundation for comparing the accuracy of two classifiers on the same prediction problem, it does have certain limitations. For example, proteins within the class 'organic acid transport' are discriminated from proteins with other BP annotations

with a relatively high ROC-score of 0.748 using the PSI-BLAST bit-score encoding of the phylogenetic profile. However, since there are only 44 yeast proteins involved in this process, this class cannot be recovered with a precision and recall that would be acceptable for guiding further experimental studies. The precision and recall are defined as

$$\text{precision} = \frac{TP}{TP+FP} \tag{5.4}$$

$$\text{recall} = \frac{TP}{TP+FN} \tag{5.5}$$

where TP is the number of true positives, FP the number of false positives and FN the number of false negatives. So to recover a single example of the 'organic acid transport' class (recall= 0.022) the precision rate is unacceptably low at 0.002. A more effective measure for assessing the potential utility of predictions for a particular GO class is the point at which the precision and recall rates break-even. The precision/recall break-even point, pr, is therefore used in the following section to assess which classes from the three ontologies can be predicted most usefully using phylogenetic profiles. The learning algorithm and benchmarking approach is identical to that described in Sections 5.2.1 and 5.2.2 with the linear SVMs trained using the normalized PSI-BLAST bit-score encoding of the phylogenetic profile vector. Results are also presented from the linear SVMs trained on the amino acid compositional data described in the previous section. Finally, a combined phylogenetic profile/compositional features method was developed by combining the two set of inputs into a single $93 + 24 = 117$-dimensional input vector.

5.3 Results of Predicting GO Annotations using Phylogenetic Profiles and Sequence Composition

Tables 5.4-5.6 present the results of predicting terms from the molecular function, biological process and cellular component ontologies using the data sources described in the previous section. Table 5.4 shows the MF terms that were predicted with highest accuracy according to the *pr* score with each column representing the mean precision/recall break-even point, *pr*, for linear SVMs trained on phylogenetic profiles, composition data and a combination of these two sets of inputs. The estimates of the *pr* rate and its standard error were calculated by 4-fold cross-validation with 10 random partitions of the training set (see Section 5.2.2). For a particular MF annotation, the *pr* scores recorded in bold type are significantly higher (at the 95% level) than the scores recorded in normal type. The statistical significance is calculated using a t-test on the difference between the two mean scores. Tables 5.5 and 5.6 show similar results for the biological process and cellular component ontologies, respectively.

The most striking feature of Table 5.4 are the scores for the term RNA-directed DNA polymerase activity, which is predicted with very low *pr*-score using either sequence composition or phylogenetic profiles in isolation, but is predicted very accurately using the combined vector. Other terms that are predicted with far higher accuracy using a combination of the two data sources include protein kinase activity, RNA helicase activity, ligase activity (phosphoric esters), GTPase activity and nuclease activity. These are relatively small classes and it seems intuitively reasonable that these types of proteins could be co-inherited in related organisms.

The majority of the molecular functions that are predicted with highest precision/recall break-even rates are catalytic activities with the only exceptions being the

molecular function	protein number	*pr* profiles	*pr* sequence	*pr* combined
DNA polymerase (RNA-directed)	51	8.75 ± 0.98	2.35 ± 0.44	**78.42 ± 8.17**
protein kinase	126	48.71 ± 1.26	15.68 ± 2.29	**77.64 ± 0.45**
catalytic	1794	67.24 ± 0.25	**70.55 ± 0.10**	**70.35 ± 0.21**
structural c'tuent ribosome	213	45.20 ± 0.43	66.61 ± 0.23	**68.84 ± 0.41**
RNA helicase	39	40.13 ± 2.79	10.29 ± 1.22	**64.00 ± 0.98**
transporter	409	35.33 ± 1.13	**61.38 ± 0.14**	58.90 ± 0.31
structural molecule	328	39.68 ± 0.43	**50.58 ± 0.21**	**48.86 ± 0.51**
kinase	198	25.82 ± 0.90	11.68 ± 0.73	**49.99 ± 1.61**
binding	881	43.30 ± 1.32	42.17 ± 0.61	**49.13 ± 0.22**
ligase (phosphoric esters)	39	26.19 ± 0.70	17.00 ± 1.34	**47.09 ± 1.37**
helicase	77	18.68 ± 1.73	18.67 ± 2.23	**43.41 ± 0.91**
nuclease	134	8.68 ± 0.47	10.17 ± 0.32	**40.85 ± 4.35**
chaperone	84	27.23 ± 0.68	6.99 ± 0.49	**39.91 ± 1.14**
hydrolase (glycosyl bonds)	40	5.59 ± 0.98	20.97 ± 1.24	**39.69 ± 3.14**
oxidoreductase	236	**37.30 ± 0.64**	29.18 ± 0.41	**39.17 ± 0.66**
hydrolase (acid anhydrides)	221	27.55 ± 1.51	20.30 ± 0.76	**37.78 ± 1.47**
P-P-bond-hydrolysis-driven transporter	57	15.58 ± 1.09	4.83 ± 0.86	**35.78 ± 1.82**
ATPase, coupled	101	21.41 ± 0.49	12.38 ± 1.42	**33.74 ± 1.67**
DNA polymerase (DNA-directed)	69	5.18 ± 0.89	2.13 ± 0.52	**32.63 ± 7.48**
nucleic acid binding	552	20.95 ± 0.41	26.66 ± 1.30	**32.16 ± 1.44**
transferase (phosphorus-containing groups)	325	18.89 ± 1.14	11.38 ± 0.75	**31.47 ± 2.27**
transcription regulator	310	21.26 ± 0.27	29.49 ± 0.38	**29.66 ± 0.52**
hydrolase	685	22.58 ± 0.94	25.22 ± 0.51	**28.07 ± 0.53**
electrochemical potential -driven transporter	30	5.42 ± 0.82	4.93 ± 0.85	27.97 ± 1.75
RNA binding	306	14.34 ± 0.14	13.07 ± 0.35	**27.44 ± 1.23**
transferase	633	23.07 ± 0.62	23.67 ± 0.60	24.96 ± 0.73
GTPase	47	3.19 ± 0.53	2.03 ± 0.06	**24.07 ± 4.44**

Table 5.4: Terms from the molecular function ontology that are predicted with highest precision/recall break-even point, *pr*, using phylogenetic profiles, sequence composition and a combination of these two data sources. 'Protein number' refers to the number of proteins in yeast that are annotated with the particular GO term. The word 'activity' has been removed from the MF term names to reduce the length of each description.

biological process	protein number	*pr* profiles	*pr* sequence	*pr* combined
protein amino acid phosphorylation	93	**43.26 ± 3.62**	8.20 ± 0.83	**54.19 ± 0.86**
biosynthesis	840	50.96 ± 0.17	47.25 ± 0.30	**52.81 ± 0.24**
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	1332	45.91 ± 0.51	47.57 ± 0.79	**51.19 ± 0.40**
protein biosynthesis	450	44.42 ± 0.31	42.61 ± 0.26	**49.82 ± 0.10**
transport	806	39.54 ± 0.46	48.17 ± 0.18	**49.48 ± 0.22**
protein metabolism	1127	46.89 ± 0.14	36.69 ± 0.27	**47.40 ± 0.27**
organic acid metabolism	257	**46.76 ± 0.28**	31.78 ± 0.46	**46.54 ± 0.34**
amino acid and derivative metabolism	187	**41.73 ± 0.59**	25.94 ± 0.80	**44.47 ± 0.59**
DNA transposition	106	**42.39 ± 5.18**	0.76 ± 0.02	**33.14 ± 7.23**
protein modification	376	**27.79 ± 2.74**	15.99 ± 1.06	**32.93 ± 1.25**
carbohydrate catabolism	39	**30.48 ± 2.52**	10.60 ± 1.40	**31.28 ± 1.76**
RNA metabolism	407	25.06 ± 0.47	19.29 ± 0.88	**30.91 ± 0.43**
glucose metabolism	54	**30.82 ± 0.90**	7.60 ± 1.60	**30.79 ± 1.76**
ribosome biogenesis and assembly	216	22.59 ± 0.41	24.88 ± 0.54	**28.32 ± 0.37**
RNA modification	76	**22.37 ± 0.66**	4.77 ± 0.61	**26.83 ± 1.16**
monosaccharide metabolism	84	**23.53 ± 0.58**	11.03 ± 0.69	**26.80 ± 1.56**
alcohol metabolism	144	**22.57 ± 0.76**	19.65 ± 0.62	**26.64 ± 0.99**
carbohydrate metabolism	65	**22.54 ± 1.64**	10.59 ± 0.64	**25.99 ± 1.35**
sterol metabolism	34	**20.59 ± 1.47**	4.06 ± 0.74	**25.05 ± 1.47**
transcription	445	18.55 ± 0.53	22.14 ± 0.46	**24.98 ± 0.42**
rRNA metabolism	154	18.84 ± 0.67	19.24 ± 0.65	**24.85 ± 0.48**
phosphorus metabolism	165	**20.08 ± 3.56**	4.77 ± 0.56	**24.26 ± 2.88**
protein folding	43	**21.22 ± 1.67**	3.71 ± 0.43	**24.13 ± 0.82**
tRNA metabolism	87	**20.55 ± 1.02**	5.17 ± 0.50	**23.74 ± 0.78**
vesicle-mediated transport	247	18.39 ± 0.43	15.41 ± 1.04	**23.69 ± 0.31**

Table 5.5: Terms from the biological process ontology that are predicted with highest precision/recall break-even point, *pr*, using phylogenetic profiles, sequence composition and a combination of these two data sources. 'Protein number' refers to the number of proteins in yeast that are annotated with the particular GO term.

cellular component	protein number	*pr* profiles	*pr* sequence	*pr* combined
cytosolic ribosome	150	56.57 ± 0.72	61.99 ± 0.43	**65.17 ± 0.73**
ribosome	256	46.53 ± 0.35	58.80 ± 0.28	**61.59 ± 0.34**
cell wall	94	14.51 ± 0.71	**58.56 ± 0.50**	52.13 ± 0.62
nucleus	1825	47.47 ± 0.24	**55.57 ± 0.10**	**55.98 ± 0.13**
cytosol	349	48.83 ± 0.23	41.07 ± 0.24	**51.06 ± 0.44**
ribonucleoprotein complex	396	40.76 ± 0.24	**48.92 ± 0.41**	**50.57 ± 0.41**
membrane	848	31.52 ± 0.98	**49.83 ± 0.16**	48.77 ± 0.16
retrotransposon nucleocapsid	94	10.23 ± 5.35	0.76 ± 0.03	**45.84 ± 3.62**
mitochondrion	654	39.23 ± 0.15	41.42 ± 0.73	**44.92 ± 0.20**
organellar ribosome	77	35.28 ± 0.52	20.31 ± 0.60	**43.35 ± 0.98**
plasma membrane	213	24.47 ± 1.97	33.94 ± 0.61	**39.74 ± 0.40**
organellar large ribosomal subunit	43	24.16 ± 0.81	10.76 ± 1.08	**34.06 ± 0.49**
mitochondrial matrix	139	27.61 ± 0.38	19.42 ± 0.29	**33.02 ± 0.59**
eukaryotic 43S preinitiation complex	68	**28.33 ± 1.34**	21.25 ± 1.12	**32.31 ± 1.17**
small ribosomal subunit	94	25.42 ± 0.74	24.10 ± 0.67	**32.16 ± 0.49**
endoplasmic reticulum	352	8.38 ± 0.39	**26.06 ± 0.31**	**25.14 ± 0.53**
nucleolus	192	9.44 ± 0.46	23.82 ± 0.39	**26.05 ± 0.52**
organellar small ribosome subunit	33	**22.35 ± 1.60**	3.41 ± 0.35	**24.38 ± 1.06**
mitochondrial membrane	179	21.03 ± 0.57	18.47 ± 1.27	23.53 ± 0.68
proteasome complex	45	2.62 ± 0.38	1.88 ± 0.11	**22.23 ± 2.38**
nucleoplasm	264	**20.93 ± 0.30**	13.69 ± 0.52	**19.46 ± 0.69**
small nuclear ribonucleoprotein complex	60	11.46 ± 0.69	10.83 ± 0.88	**20.00 ± 0.93**
membrane coat	34	11.77 ± 1.27	1.95 ± 0.05	**19.96 ± 0.67**
RNA polymerase complex	31	**17.74 ± 1.27**	2.76 ± 0.14	**19.39 ± 2.28**
integral to membrane	204	5.27 ± 0.33	**14.75 ± 0.32**	**15.54 ± 0.32**
spliceosome complex	63	6.75 ± 0.67	3.52 ± 0.57	**15.09 ± 0.56**
small nucleolar ribonucleoprotein complex	31	3.82 ± 0.55	5.22 ± 1.11	**13.71 ± 0.95**

Table 5.6: Terms from the cellular component ontology that are predicted with highest precision/recall break-even point, *pr*, using phylogenetic profiles, sequence composition and a combination of these two data sources. 'Protein number' refers to the number of proteins in yeast that are annotated with the particular GO term.

very general classes of binding, transporter and structural molecule activity. These and other large, heterogeneous functional classes such as catalytic activity are predicted with highest accuracy using sequence composition data alone. It might be expected that the SVMs trained on phylogenetic profile data would produce poor accuracy on these classes, as the profiles are unlikely to be separable using a linear decision function. Preliminary investigations suggest that a significant improvement in accuracy can be achieved by the use of non-linear kernel functions for these classification problems. It is likely that the large functional classes such as catalytic activity and transporter activity are predicted relatively accurately using sequence composition data alone because they are characterized by strong structural signals. For example, enzymes are typically water-soluble, contain low predicted disorder composition (previous chapter) and are of mixed α/β composition (Martin et al., 1998).

The results for the molecular function classes in Table 5.4 are promising in terms of accuracy and potential biological significance. However, it is necessary to exercise some caution. Although the homology thresholds used to benchmark the performance of the classifiers are fairly strict, molecular function is often determined by a small number of conserved residues. It may therefore be the case that sequence motif or tuned hidden Markov models (Hulo et al., 2004; Bateman et al., 2004) will allow identification of proteins within these families with even greater accuracy. A method directed specifically at identifying novel kinase families, for example, would necessarily be assessed by using more stringent homology thresholds (e.g. by leave-one-out cross-validation using each known family of protein kinase) but this is beyond the scope of this chapter.

The results from Table 5.5 indicate that phylogenetic profiles can be used to predict biological process classes more accurately than sequence composition, and that the combined feature vector typically leads to a slight increase in prediction

accuracy. The greatest exception to this trend occurs in the prediction of proteins involved in DNA transposition. As described in Chapter 4, the Ty-elements are thought to have evolved from a retroviral infection which became fixed in the yeast genome. These proteins are therefore unique to *Saccharomyces* and the fission yeast *Schizosaccharomyces pombe* but are distantly related to transposable elements in *Drosophila*, mouse, rat, human and maize. These proteins are therefore likely to have fairly distinct phylogenetic profiles. Supplementing the profile data with information on sequence composition perhaps leads to a dilution of this strong signal and greater confusion between the DNA transposition class and other nuclear proteins.

The majority of the process classes that are predicted accurately in Table 5.5 are metabolic with the only exceptions being the DNA transposition and ribosome biogenesis and assembly classes. This also appears to be biologically reasonable as these pathways perform specific biochemical functions that would appear to be independent of context. Conversely, processes such as cell signalling are more likely to be adapted to the organisms' particular environment. The other requirement for these process classes to be identified accurately using phylogenetic profiles is that they have patterns of inheritance that differ from other functional classes. This property must also apply to pathways such as yeast alcohol and sterol metabolism, which are only shared by a subset of the organisms in the phylogenetic comparison.

It is difficult to compare results from the BP ontology with those of previous studies on *Saccharomyces cerevisiae*, as these appear not to have structured the cross-validation so that homologous proteins do not occur in both test and training sets, and have predicted MIPS functional categories rather than those defined using the Gene Ontology (Pavlidis et al., 2002; Vert, 2002; Lanckriet et al., 2004). However, there is significant overlap in the classes that are predicted with highest accuracy, such as chaperones, transporters, and ribosomal proteins.

A general trend in the results for the cellular component ontology (Table 5.6) which differs from the other two ontologies is that sequence composition contributes more to the accurate prediction of sub-cellular location than phylogenetic profiles. This fact is also reflected in the literature on sub-cellular location prediction which has shown that locales such as the nucleus, mitochondrion and extra-cellular can be identified using measures based on sequence composition (Nakashima and Nishikawa, 1992; Park and Kanehisa, 2003). The large cellular compartments such as the nucleus and mitochondrion also contain large and heterogeneous sets of proteins involved in numerous biochemical pathways, and are likely to be predicted more accurately using a non-linear kernel function.

Organelles such as the mitochondrion and plant chloroplasts are believed to have originated from prokaryote cells that became fixed in an ancestor of eukaryotes (Schwartz and Dayhoff, 1978). It might therefore be expected that these proteins have a clear pattern in their phylogenetic profiles. However, the organelles became part of eukaryote cells in the distant evolutionary past, and many of the proteins that were present in the prokaryote progenitor of the mitochondrion appear to have been translocated to other cellular compartments in the course of evolution (Marcotte et al., 2000). As a result the phylogenetic signal for this organelle has largely been diluted to the extent that the mitochondrial plasmid genome now contains only around 20 genes encoding large membrane proteins, which are retained in the mitochondrion as they cannot be transported easily from the endoplasmic reticulum (Dwight et al., 2002).

The most promising aspect of Table 5.6 is that the combination of sequence composition data with phylogenetic profiles can be used to identify protein complexes such as the ribosome, eukaryotic 43S preinitiation complex and the spliceosome complex. This is plausible from a biological perspective, as it would be expected that proteins that form part of a functional complex would be co-inherited as intact mod-

ules during the course of evolution. The relatively high accuracy of the predictions for small classes such as the large ribosomal subunit also suggests that complexes containing even fewer proteins may be identifiable with phylogenetic profiles. The limitation of using phylogenetic profile data in isolation is that other unrelated complexes or biological pathways may show a similar pattern of inheritance across complete genomes. The combination of phylogenetic profiles with another complementary data source may therefore allow these complexes to be discriminated from each other. Although sequence composition is a very simple means of encoding the amino acid sequence, there are several other sources of information that are likely to be more effective in identifying protein complexes or pathways. These are described in the following section which also summarizes the results and suggests directions for future work in this area.

5.4 Discussion

The phylogenetic profile is another example of using the inherent modularity of biological systems to investigate protein function. It appears that Nature's parsimonious adaptation of existing biological units extends in scale from short sequence motifs to entire complexes and pathways. The classes that are predicted with highest accuracy therefore represent the biochemical functions, pathways or complexes that tend to be inherited as intact modules between different organisms.

The results for the molecular function ontology suggest that phylogenetic profiles and sequence composition data may provide another means for assigning some types of biochemical activity. The accurate prediction of 'protein kinase activity' may be of great biological significance as kinases are one of the most important classes of regulatory proteins in the cell. However, as described in the introduction of this chapter, molecular function is often determined by a small number of conserved

residues, and the various families of protein kinase tend to contain highly conserved catalytic sites (Bossemeyer, 1995).

Although phylogenetic profiles may contain orthogonal information, which could prove useful for improving the assignment of molecular function, it is likely that this aspect of protein function will continue to be predicted most accurately using traditional homology-based approaches and structural models (Pazos and Sternberg, 2004). The main advantage of phylogenetic profiles lies in inferring higher-level functional properties such as protein-protein interactions, which are difficult to detect both experimentally (Jansen et al., 2003) and using sequence/structure-based methods (Wodak and Mendez, 2004).

The results from Tables 5.5 and 5.6 indicate that phylogenetic profiles are most effective for small, specific biological processes or protein complexes, and this may be a more promising direction for future research. This is consistent with the most well-established methods for assigning function which do not attempt to learn decision rules for entire functional classes but use either close co-inheritance or other evolutionary signals (described in greater detail in the following section) to establish functional links between proteins (Marcotte et al., 1999b; Huynen et al., 2003; von Mering et al., 2005). These links can then be used to predict shared protein functions with low coverage but much higher precision.

An advantage of using pair-wise functional links is that they represent either a direct physical interaction or close proximity within a particular pathway rather than a subjective functional label such as GO or KEGG annotations (Kanehisa et al., 2004). It has even been argued that physical interactions, common evolutionary origin, co-regulation and other biological properties of the protein provide a more realistic definition of protein function than artificial schemes such as the Gene Ontology (Fraser and Marcotte, 2004). A more effective use of supervised learning

algorithms may therefore attempt to predict whether putative links represent a true functional relationship, and whether the link represents a transient physical interaction, membership of the same complex, or participation in the same biochemical pathway.

5.4.1 Future Work

The scope for extending the work presented in this chapter is enormous, since the investigation of the 'function' carried out by gene products is the fundamental objective of a large proportion of the research in experimental and computational biochemistry. There is now a large corpus of experimental data for investigating biological systems that comes from genome-wide experiments such as microarrays (Brazma and Vilo, 2000; Brown et al., 2000), 2-hybrid screens and pull-down assays for detecting protein-protein interactions and protein complexes (Gavin et al., 2002; Uetz et al., 2000), and systematic gene 'knock-outs' (Winzeler, 1999). There are also several publicly-available databases containing information that is mined from the scientific literature, such as the Database of Interacting Proteins (DIP) (Xenarios et al., 2000).

The major advantages of the algorithms described in this chapter is that they utilize data which can be derived directly from the completed genome sequence, may be extensible to other eukaryote organisms and do not require further expensive experimental studies of the organism. The discussion of the potential extensions of this work is therefore restricted to improving the prediction of protein function using either amino acid or DNA sequences. These extensions involve improving the existing data representations, incorporating other complementary sources of information from the protein or DNA sequence, and combining the data sources more effectively. The following section describes potential improvements and extensions

of the existing information resources, and the subsequent section suggests machine learning techniques that could be used to combine these data sources into an overall prediction of protein function.

Improved and Extended Data Representations for Inferring Protein Function

There are several potential improvements that could be made to the representation of the phylogenetic profile described in this chapter. The most successful use of phylogenetic profiles, to be described in the literature, has been in detecting very similar patterns of inheritance between proteins that typically consist of a single domain (Huynen et al., 2003; von Mering et al., 2005). The difficulty in representing multi-domain proteins with phylogenetic profiles is that information on which domain is aligned with a homologue in the comparison genome is not contained within the profile vector. For example, a two-domain protein consisting of one domain of prokaryote origin and another domain that is unique to eukaryotes may be recorded as being inherited across all genomes in the phylogenetic profile vector. It may therefore be more effective to use the patterns of co-inheritance at the level of single domains. This could be achieved by using fold recognition methods (McGuffin and Jones, 2003) or curated databases of domains such as SMART or PFAM (Letunic et al., 2004; Bateman et al., 2004). The state-of-the-art cluster of orthologous group (COG) database, which curates groups of proteins with similar patterns of orthology, has now been extended to several eukaryote organisms including yeast, and could potentially be used to infer the orthology patterns of single domains (Tatusov et al., 2003).

The other data sources that are likely to be most relevant to improving prediction of cellular component and biological process are signal peptide predictions,

which encode location targeting signals (Bendtsen et al., 2004) and transmembrane helix predictions (Kall et al., 2004). The sequence composition data could also be supplemented with simple physical properties of the protein such as the net charge, hydrophobicity and isoelectric point, which have been shown to be effective for predicting some BP classes in the Human genome (Jensen et al., 2003). It is also possible that incorporating predictions of other active sites such as metal binding sites (Sodhi et al., 2004a,b) and phosphorylation sites (Iakoucheva et al., 2004) could improve predictions of biological process or cellular component.

Most of the potential sources of data that can be derived from the amino acid sequence have been described previously but there is one other potential candidate for identifying protein-protein interactions. Gene fusions are proteins that are observed as a single, fused protein in one organism but as two separate proteins in another organism (Marcotte et al., 1999a). It has been postulated that one evolutionary advantage of fusing two enzymes that are nearby in a biochemical pathway is that the co-regulation of protein concentration is automatically achieved, provided that the proteins have a 1:1 stochiometry. Another advantage is that the effective concentration of the fused catalyst is also greater than that of the separate components, so the same task can be performed with the translation of fewer proteins. The participation of two proteins in a fusion event therefore implies either a direct physical interaction or involvement in the same pathway.

A recent paper by Marcotte and Marcotte (2002) showed that it is possible to calculate a measure of the confidence that a fusion event represents a true physical interaction. This was based on a very simple statistical model based on the frequency of the component domains and the fused protein. Some preliminary work that we have carried out in this area suggests that the number of paralogues of the two domains and several properties of the fused protein can be used to discriminate fusion events that indicate a functional link from those that do not. The other

finding is that remote homology detection using profile-based searching algorithms can be used to identify more evolutionarily distant fusion events without greatly compromising specificity.

Although this work has concentrated on using amino acid sequences to predict protein function, there is likely to be significant information contained within the *cis*-regulatory regions which mediate control of transcription. Co-regulation is also suggested by conserved gene order or *synteny* of sets of genes across different organisms (Huynen et al., 2003). Although co-regulation is not imposed at the genome level in higher eukaryotes (Blumenthal et al., 2002), gene order is important in bacteria since co-expression is usually achieved by locating genes on the same operon. It has been shown that a functional link can be assigned to two eukaryote proteins that have orthologues in prokaryotic genomes with conserved chromosomal locations (Huynen et al., 2000). In addition to these indirect means of identifying co-expressed genes, it is also possible to detect specific transcription factor binding sites using resources such as transfac database (Matys et al., 2003). The advent of microarray technology is now also allowing investigations of complex patterns of gene regulation involving multiple interactions between promoter, repressor and non-specific transcription factors (Beer and Tavazoie, 2004), and in the future these techniques are likely to prove extremely powerful for inferring and characterizing protein function.

Machine Learning Techniques for Combining Data Sources

The integration of data from numerous sources into an overall prediction is an emerging trend in functional genomics (Marcotte et al., 1999b; Jansen et al., 2003; Lanckriet et al., 2004) but combining different data sources is only useful if they encode information that is complementary (Duda et al., 2000). The results of this chapter suggest that a combination of phylogenetic profiles and sequence-derived features

improves the accuracy of predicting most protein functional classes. The previous section has also described several other potential sources of data that have proven useful in inferring protein function. However, simply concatenating several sources of information into a single feature vector may not be an optimal approach for combining heterogeneous data sources. In general, the concatenated feature vector will have high dimensionality and may make the learning system more prone to overfitting (Vapnik, 1998). In the case of support vector machines, the use of a single feature vector also reduces the flexibility in choosing the kernel function for the particular learning task and prevents using the kernel to encode discrete structures such as strings and graphs (Lodhi et al., 2000; Kondor and Lafferty, 2002). This is a severe restriction as this property is likely to be crucial for encoding several of the data sources described in the previous section.

The challenge from a machine learning perspective is how to integrate these sources of data into an improved prediction of either protein-protein interactions or functional classification. Another advantage of kernel methods is that simple arithmetic combinations of valid kernel functions (which generate positive, semi-definite kernel matrices) also result in valid kernels. Some examples of the operations for combining kernels are shown below

$$K(\mathbf{x},\mathbf{z}) = aK_1(\mathbf{x},\mathbf{z}) + bK_2(\mathbf{x},\mathbf{z}) \tag{5.6}$$

$$K(\mathbf{x},\mathbf{z}) = K_1(\mathbf{x},\mathbf{z})K_2(\mathbf{x},\mathbf{z}) \tag{5.7}$$

$$K(\mathbf{x},\mathbf{z}) = f(\mathbf{x})f(\mathbf{z}) \tag{5.8}$$

where $K_1(\mathbf{x},\mathbf{z})$ and $K_2(\mathbf{x},\mathbf{z})$ are two kernel functions that obey Mercer's condition (see Introduction), a and b are two real constants, and $f(\cdot)$ is a real-valued function.

This property has led to the development of an algorithm for obtaining optimal

linear combinations of kernel functions generated from several data sources (Lanckriet et al., 2004). The algorithm, which is based on a semi-definite programming, has been shown to provide accuracies that are significantly higher than using any of the individual data sources in isolation. The goal of 'learning the kernel function' is an active topic of research in machine learning but whether this will eventually provide improvements over other techniques for combining information from different data sources is an open question.

Despite the advantages of kernels methods described in this section, the results of chapters 2, 3 suggest that SVMs do not perform with greater accuracy than neural networks on the disorder or secondary structure prediction problems. The subsequent chapter discusses some of the reasons for this, and also summarizes the most important biological contributions made by this thesis.

Chapter 6

Discussion

This chapter is divided into two halves which summarize the major contributions of this thesis to biology and to the practical use of supervised learning algorithms to problems in bioinformatics.

6.1 Biological Discoveries

The first chapter indicated that a consensus could improve the accuracy of secondary structure prediction, and that existing methods could therefore be improved further. However, most of the increase in accuracy occurs in predicting the ends of regular structural elements (α-helix and β-sheet), and it is questionable whether this improvement would contribute a great deal to either three-dimensional structure prediction methods or human experts. The termini of the regular structural elements are also less well defined, and the increased accuracy may arise from the classification system learning properties of the DSSP secondary structure assignment algorithm (Kabsch and Sander, 1983) rather than a genuine physical property of the protein.

In Chapter 2, the fact that improvements in the accuracy of secondary structure prediction have plateaued in recent years is presented as evidence that local sequence-based prediction methods are approaching an upper limit on their potential accuracy. It is suggested that the limitation arises as result of secondary structure not being fully conserved between homologous proteins (Rost, 2001a) and the dependence of secondary structure on non-local interactions. Significant increases in accuracy are therefore likely to involve incorporating longer-range interactions into a complete three-dimensional model of protein structure (Meiler and Baker, 2003).

Although this thesis has provided some insight into the secondary structure prediction problem, the most novel and important contributions have been in under-

standing the functions and origins of disorder in eukaryote genomes. The observation that direct use of the amino acid sequence improves accuracy of disorder prediction over methods based on sequence composition indicates that disorder is not a simple structural feature based on either the net charge or hydrophobicity. Another important observation is that information on the conservation of the amino acid sequence can be used to improve slightly the accuracy of disorder prediction. The fact that this improvement is far smaller than in secondary structure prediction suggests that the difference in conservation between disordered and ordered residues is not as great as between ordered coil residues and the regular secondary structure elements. Evolutionary information improves the accuracy of secondary structure prediction by encoding long-range constraints on the local structure (see Chapter 2). Since, by definition, disordered residues are not subject to strong non-local interactions with other parts of the protein, most of the information required to identify regions of native disorder is likely to be contained in the primary sequence, as shown by Chapter 3.

I speculate that some disordered regions are highly conserved (e.g. those in molecular recognition sites) and others, such as those that are present in flexible domain linkers and other entropic chains, are less well-conserved. The conservation of disordered regions has been studied superficially (Brown et al., 2002) but is an area that is in need of further study. The problem is complicated by the low information content of many disordered sequences, which is a result of many disordered structures evolving *via* repeat expansion (Tompa, 2003). This could potentially be overcome by using techniques from computational linguistics (Durbin et al., 1998).

The improved accuracy of the DISOPRED2 classifier allowed the identification of disordered residues and particularly long (> 30 residue) disordered regions with very low false positive rates[1]. The study described in Chapter 3 of the frequency

[1]The per residue false positive disorder prediction rate was 3.2% and the per chain false prediction

of disorder in the three kingdoms of life therefore has the advantage over previous studies of being based on a more accurate prediction method with a far lower disorder over-prediction rate. The results indicate that disorder is a common feature of eukaryote proteins but is far less common in prokaryotes. This is consistent with experimental studies, which have shown that dynamic flexibility of the protein structure is more often associated with eukaryote protein functions (Tompa, 2003), and is one explanation for the greater difficulty of crystallizing many eukaryote proteins (Goh et al., 2004). The initial analysis of the association between domain cuts (Chandonia et al., 2004) and predicted disorder, described in Chapter 4, is consistent with dynamic flexibility being a property of many domain linkers (Dyson and Wright, 2002).

The second half of chapter 4 used the proteome of the model organism *Saccharomyces cerevisiae* to investigate the function of long segments of predicted disorder in eukaryote cells. The results reinforce the experimental evidence for native disorder being involved in DNA-binding, signal transduction and modification of the cytoskeleton (Wright and Dyson, 1999; Iakoucheva et al., 2002), and also suggest other roles for disorder in eukaryote cells. These roles include cellular morphogenesis, kinase activity and DNA packaging. The results from the cellular component ontology also suggested that disordered proteins are located in cellular compartments that provide some protection from proteolysis such as the nucleus and cell cortex. The absence of cellular compartments and highly-regulated degradation pathways in prokaryotes is postulated as one reason for the lower frequency of disorder in these organisms.

The results in Chapter 4 also suggest pathways of clinical importance that involve proteins containing a large number of long, disordered segments. These biological processes, which include transcription regulation, DNA repair, cell cycle and epige-

of long disordered regions was less than 0.1%

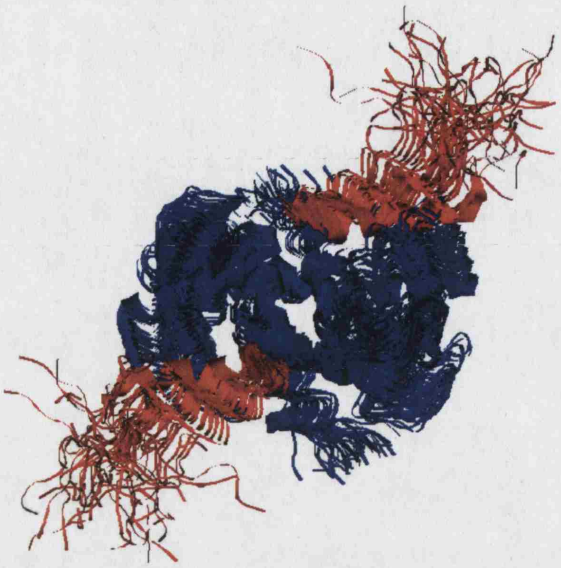

Figure 6.1: NMR structure of the C-terminal negative regulatory domain of *p53* in a complex with Ca^{2+}-bound S100B. The predicted regions of disorder (coloured in red) are modelled by 40 isoforms and appear to be highly flexible in solution.

netic control of gene expression provide further clues as to the causal role of disorder in many forms of cancer (Iakoucheva et al., 2002). The success of DISOPRED2 in detecting disorder in proteins within these clinically relevant pathways is demonstrated by Figure 6.1, which shows accurate identification of disordered regions within the well-known tumour suppressor protein *p53*.

The *jackknifed* cross-validation experiment for predicting protein function, described in Chapter 5, showed that non-homologous proteins within certain pathways or cellular locations show similar patterns of inheritance across complete genomes. The other contribution of the final research chapter is the demonstration that phylogenetic profile and sequence composition data can be used to predict several more specific functional classes than those that have been identified previously. The results indicate that specific sub-cellular locations such as the mitochondrial matrix

and complexes such as RNA polymerase can be predicted accurately using these sources of data.

The inclusion of other higher-order properties of the amino acid sequence such as domain composition (Mott et al., 2002), sequence motifs (Hulo et al., 2004) and predictions of transmembrane regions and signal peptides (Kall et al., 2004), is likely to lead to further improvements in prediction accuracy and may allow the identification of even more specific cellular component annotations. The advantage of these sources of information is that they can be generated directly from a complete proteome, and can therefore be applied easily to newly-sequenced organisms. Incorporating these data sources into an overall prediction of protein complex or perhaps biochemical pathway is likely to be the direction of future research in this area.

The following section discusses some of the implications of this thesis for applications of supervised learning algorithms to problems in bioinformatics.

6.2 Application of Kernel-Based Machine Learning to Problems in Bioinformatics

Chapters 2 and 3 indicate that SVMs do not necessarily outperform neural networks on pattern recognition problems in bioinformatics. There are two potential reasons for this being the case. Firstly, the learning bias used by the support vector machine to select a model of the training data may not provide better generalization than the learning bias used to train the feed-forward neural network. The second potential cause is that the projection of the data to a higher dimensional feature space, achieved by the kernel function in SVMs, may not provide a better separation of the classes than the layer of hidden units used to solve non-linear problems with feed-forward neural networks.

A characteristic of most pattern recognition problems in bioinformatics, including those dealt with in this thesis, is high empirical error, and it is not clear that the SVM's learning bias provides good generalization in these cases. The probably approximate correct (PAC) justification for margin maximization does not apply to problems with high empirical error (Cristianini and Shawe-Taylor, 2000), and the form of the SVM solution suggests poor generalization both intuitively and from the perspective of sample compression. The intuitive explanation can be visualized by splitting the SVM decision function into the separate contributions from bounded ($\alpha_i = C$) and ($0 < \alpha_i < C$) unbounded support vectors.

$$f(\mathbf{x}) = \sum_{i:\gamma_i=1} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + C \sum_{i:\gamma_i<1} y_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{6.1}$$

where γ_i is the geometric distance from the separating hyperplane for example $\mathbf{x}_i$.

This means that misclassified examples ($\gamma_i < 0$) appear in the solution but correctly classified examples outside the geometric margin do not. So using the example of the disorder prediction problem, the most easily recognized ordered residues do not appear in the SVM solution. However, disordered residues that are falsely labelled as ordered because they are stabilized by a protein interaction appear in the SVM decision function as representatives of the ordered class. It therefore seems reasonable that greater accuracy would arise in these problems by using the "best" examples of each class in the decision function. Some preliminary results from applying linear and kernel Fisher discriminants (Mika et al., 1999) to the function prediction problem suggest that these solutions, which are based on all of the training examples, provide better accuracy than SVMs for this problem.

The theoretical justification comes from the *minimum description length* (MDL) principle, which views the process of learning as encoding the training data with

the most efficient possible description. The description length $K(h, \mathcal{D})$ is therefore given by the sum of the model's algorithmic complexity, $K(h)$, and the description of the training data with respect to that model $K(\mathcal{D}$ using $h)$ (Duda et al., 2000)

$$K(h, \mathcal{D}) = K(h) + K(\mathcal{D} \text{ using } h) \tag{6.2}$$

The optimal model h^* according to the MDL criterion gives the minimum description length $h^* = \arg\min_h K(h, \mathcal{D})$. Although the MDL is often difficult to compute in practice[2], the complexity of a non-linear SVM classifier resulting from a problem with high empirical error suggests a high MDL and therefore poor generalization.

The overall conclusion of this section is that when applying classification algorithms as a 'black box', the use of a more sophisticated learning algorithm does not guarantee greater accuracy. This is a demonstration of perhaps the most fundamental result from theoretical machine learning, popularly known as the 'No Free Lunch Theorem'. The 'No Free Lunch Theorem' states that when uniformly averaging over *all possible* target functions F, the expected off-training set error or generalization is identical for any two learning algorithms (Duda et al., 2000). Consequently, if there is no prior information on the target function to be learned, there is no theoretical reason for preferring one learning algorithm over another. This is the case for all learning algorithms including SVMs, neural networks, RBF networks, nearest-neighbour classifiers, or even a random guess classifier. The main reason for some machine learning techniques becoming established as standard tools in a diverse array of fields is that there are similarities in the target functions that must be learned between the various pattern recognition problems.

Experience of many supervised learning problems in bioinformatics, and other

[2]For example, the SVM model could be described by a subset of the support vectors (Section 2.4.1)

problem domains such as text mining, indicates that the step that is most crucial to the success of the learning algorithm is the representation of the feature vector used to train the classifier. This representation can be improved by using prior knowledge of the problem to choose inputs that best represent the classification problem or by projecting the input data to a space that will allow good class separation. The first approach is a common source of improvement for methods in bioinformatics, and can be used in conjunction with any learning algorithm. One advantage of kernel methods such as Support Vector Machines is that the kernel function provides a great deal of flexibility in choosing a projection of the data to a high-dimensional feature spaces. This can be used to encode biological knowledge (Vert, 2002) or to classify non-Euclidean data such as strings and graphs (Cristianini and Shawe-Taylor, 2000; Kondor and Lafferty, 2002).

In summary, the most successful methods in bioinformatics will continue to be founded on a deep understanding of the underlying biology. However, the power and flexibility of the kernel function provides another means for incorporating prior knowledge in the design of new algorithms.

Appendix A

Abbreviations

abbreviation	definition
BLAST	Basic Local Alignment Search Tool
BP	Biological Process ontology
CASP	Critical Assessment of techniques for protein Structure Prediction
CC	Cellular Component ontology
CD	Circular Dichroism spectroscopy
DAG	Directed Acyclic Graph
DIP	Database of Interacting Proteins
DSSP	Dictionary of protein Secondary Structure Program
EcoCyc	Encyclopedia of Escherichia coli Genes and Metabolism
GO	Gene Ontology
GOLD	Genomes OnLine Database
HMM	Hidden Markov Model
KKT	Karush-Kuhn-Tucker
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
MDL	Minimum Description Length
MLP	Multi-Layer Perceptron
MF	Molecular Function ontology
ORF	Open Reading Frame
PAC	Probably Approximately Correct
PD	Proteolytic Degradation
PDB	Protein Data Bank
PSI-BLAST	Position-Specific Iterated BLAST
PSSM	Position-Specific Scoring Matrix
RBF	Radial Basis Function
SGD	*Saccharomyces* Genome Database
SOV	Segment OVerlap
SQL	Structured Query Language
SSEA	Secondary Structure Element Alignment
SVD	Singular Value Decomposition
SVM	Support Vector Machine

Table A.1: Abbreviations.

Appendix B

Definition of Scoring Schemes for Secondary Structure Prediction

There are perhaps more measures of secondary structure prediction accuracy than any other area in bioinformatics. The main reason for the proliferation of scoring schemes is that there is no obvious universal measure of how 'good' a prediction actually is. The scoring schemes are all based on the 3×3 confusion matrix M, where M_{ij} are the number of residues observed in state i and predicted in state j with $i, j \in \{H, E, C\}$.

B.1 Accuracy (Q_x) Scores

The simplest measure of accuracy is the three-state per residue accuracy or Q_3 score

$$Q_3 = 100 \cdot \frac{1}{N} \sum_{i=1}^{3} M_{ii} \tag{B.1}$$

where N is the total number of residues $N = \sum_{(i,j)} M_{ij}$. This measure of performance has some limitations since predictions that are effectively useless can still achieve a high three-state accuracy. For example, if a sequence is correctly predicted to belong to the all-alpha class, predicting one long continuous helix segment may achieve high three-state accuracy but provide little meaningful information. The unequal frequencies of the three structure classes also mean that overprediction of coil residues can lead to high Q_3 without indicating the structural segments of the protein. For this reason, there are several scores for measuring the effectiveness of recognising specific structure types. The Q_i^{obs} and Q_i^{pred} scores measure the proportion of correctly assigned residues out of the total number of residues observed to be in class i and predicted to be in class i, respectively.

$$Q_i^{obs} = 100 \cdot \frac{1}{N_i^{obs}} \sum M_{ii} \tag{B.2}$$

$$Q_i^{pred} = 100 \cdot \frac{1}{N_i^{pred}} \sum M_{ii} \tag{B.3}$$

and can be interpreted similarly to the precision and recall of binary classification. The Matthew's correlation coefficients can also be used to measure a method's success in predicting each of the three structural classes

$$C_i = \frac{p_i \cdot n_i - u_i \cdot o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)}} \tag{B.4}$$

where

$$p_i = M_{ii} \qquad n_i = \sum_{j \neq i} \sum_{k \neq i} M_{jk}$$
$$o_i = \sum_{j \neq i} M_{ji} \qquad u_i = \sum_{j \neq i} M_{ij}$$

B.2 Segment Overlap (Sov) Score

The most complicated of the performance measures for secondary structure prediction is the segment overlap or Sov-score. This score is designed to ensure that the segments outputted by the classification system have similar length distributions and significant overlap with the target protein. The latest version of sov is normalised so that scores are in the range between 0 and 100% (Zemla et al., 1999). The sov-score for a particular structure state i is

$$\text{Sov}_i = 100 \cdot \frac{1}{N_{sov,i}} \frac{\sum_{S_i} \text{minov}(s_1, s_2) + \delta(s_1, s_2)}{\text{maxov}(s_1, s_2)} \tag{B.5}$$

where $\text{minov}(s_1, s_2)$ and $\text{maxov}(s_1, s_2)$ are the minimum and maximum overlap of two segments in the target s_1 and prediction s_2 strings of secondary structure, $\delta(s_1, s_2)$ ensures normalisation and is defined as

$$\delta(s_1, s_2) = \min \left[(\text{maxov}(s_1, s_2) - \text{minov}(s_1, s_2)); \text{minov}(s_1, s_2); \frac{\text{len}(s_1)}{2}; \frac{\text{len}(s_2)}{2} \right] \tag{B.6}$$

$\text{len}(s_1)$ is the length of the segment in the target structure. The summation is taken over the set S_i of all segments in the observed structure with an overlapping partner in the prediction. The normalising factor N_{sov} is defined as

```
sequence:    -----EEEE----E--EE--EE--------EEEE-HHHEEE-
prediction:  -E---EEEEE------EE--EEEEEE----EEEE--EE-EE-
errors:       O       L   U         LLLL          LWWL
```

Table B.1: First two rows show target and prediction, with dashes representing coil structures.The four types of errors are shown in the lower row. These are over- and under-predictions, length errors andwrong predictions.

$$N_{sov,i} = \sum_{S_i} \text{len}(s_1) + \sum_{S_i'} \text{len}(s_1) \tag{B.7}$$

where the second summation is taken over the set of all segments S_i' in the observed structure that have no overlap with the prediction. The sov-scores for each structure state are combined into an overall score sov_3 by summing the unormalised scores and then normalising by

$$N_{sov3} = \sum_{i \in \{H,E,C\}} N_{sov,i} \tag{B.8}$$

As prediction accuracies have now risen to well over 75% the need for this wide variety of scoring schemes has decreased as they are all highly correlated with the Q_3-score. More recently, McGuffin and Jones (2002) showed that the specific types of error made by the classifier have the greatest bearing on fold recognition methods that incorporate secondary structure. The errors can be divided into four types (as shown in figure B.1), which emphasise the need for correct prediction of the regular helical or sheet structural elements.

	C_1	H_1	E_1
C_2	min(len(C_1), len(C_2))	0.5min(len(H_1), len(C_2))	0.5min(len(E_1), len(C_2))
H_2	0.5min(len(C_1), len(H_2))	min(len(H_1), len(H_2))	0
E_2	0.5min(len(C_1), len(E_2))	0	min(len(E_1), len(E_2))

Table B.2: **Score matrix for aligning secondary structure elements.** Each subscript refers to a sequence of secondary structure elements that are being aligned. For example, min(len(C_1), len(C_2)) gives the minimum length of two aligned coil segments.

B.3 Secondary Structure Element Alignment (SSEA) score

McGuffin and Jones (2002) simulated various types of error and showed that wrong and underpredictions are more detrimental to fold recognition than length errors. This was addressed by building a scoring scheme for secondary structure that accounts for these more deleterious types of error. The secondary structure element alignment score (SSEA) was thus suggested as a means for assessing the quality of predictions for integration into fold recognition methods. The SSEA score is calculated by aligning the true and predicted secondary structure elements using a Needleman-Wunsch algorithm for globally aligning two sequences (Needleman and Wunsch, 1970). The scoring scheme is described in (McGuffin et al., 2001) and scores the alignment between each element with the following scoring matrix.

The scores are normalized by the mean length of the two sequences and then multiplied by 100 to express the score as a percentage (McGuffin and Jones, 2002).

Appendix C

Conditions for Equivalence between Pearson Correlation Co-efficient and the Inner Product

Pearson's co-efficient of the correlation between two variables $x = x_1, \ldots, x_N$ and $y = y_1, \ldots, y_N$ is given by

$$C(x,y) = \frac{\frac{\sum x_i y_i}{N} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \tag{C.1}$$

which in the case where the means of variables x and y are normalized to zero and the variances to one gives

$$C(x,y) = \frac{1}{N}\sum x_i y_i = \frac{1}{N}\langle \mathbf{x} \cdot \mathbf{y} \rangle \tag{C.2}$$

Appendix D

Proteomes used to Generate Phylogenetic Profiles

The evolutionary information that is contained within the phylogenetic profiles is illustrated by Figure D.1, which shows the tree generated by hierarchical clustering of the vectors associated with each organism in the comparison with the yeast proteome. Table D.1 lists all of the 92 proteomes used to generate the phylogenetic profiles used in Chapter 5.

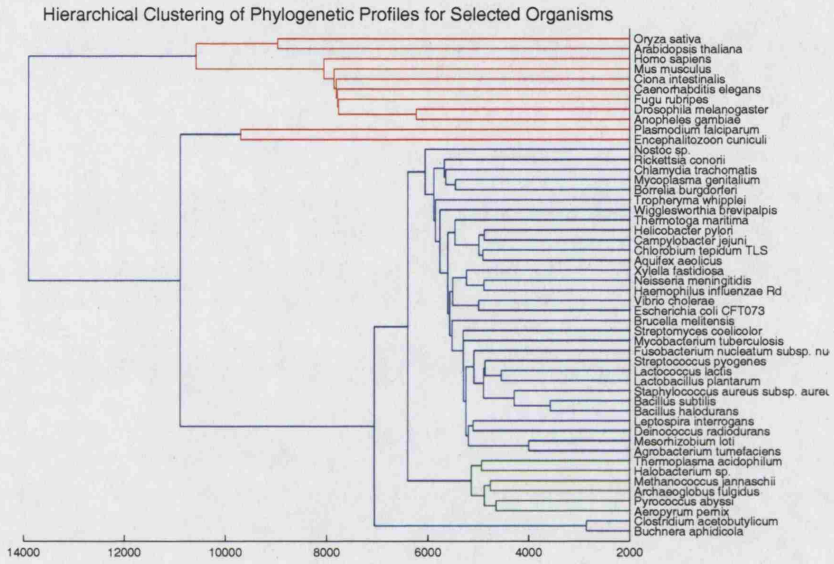

Figure D.1: Hierarchical clustering of the phylogenetic profiles for 52 of the organisms in the comparison with the yeast proteome. The clusters coloured in blue are the profiles of eubacteria, the red cluster represents eukaryote profiles, and the green cluster the profiles of archaea. The hierarchy was constructed using average linkage clustering on the Euclidean distance matrix between the phylogenetic profile vectors. The profiles were constructed using the normalized BLAST bit-score of the match between each protein in yeast and the comparison proteome.

Aeropyrum pernix	Mycoplasma penetrans
Agrobacterium tumefaciens	Mycoplasma pneumoniae
Anopheles gambiae	Mycoplasma pulmonis
Aquifex aeolicus	Neisseria meningitidis MC58
Arabidopsis thaliana	Neurospora crassa
Archaeoglobus fulgidus	Nostoc sp. PCC 7120
Bacillus anthracis str. Ames	Oceanobacillus iheyensis
Bacillus halodurans	Oryza sativa
Bifidobacterium longum NCC2705	Pasteurella multocida
Borrelia burgdorferi	Plasmodium falciparum 3D7
Bradyrhizobium japonicum	Plasmodium yoelii
Brucella melitensis	Pseudomonas aeruginosa PA01
Buchnera aphidicola (Baizongia pistaciae)	Pseudomonas syringae DC3000
Caenorhabditis briggsae	Pyrobaculum aerophilum
Caenorhabditis elegans	Pyrococcus abyssi
Campylobacter jejuni	Pyrococcus furiosus DSM 3638
Caulobacter crescentus CB15	Pyrococcus horikoshii
Chlamydia trachomatis	Ralstonia solanacearum
Chlorobium tepidum TLS	Rickettsia conorii
Ciona intestinalis	Saccharomyces cerevisiae
Clostridium acetobutylicum	Salmonella enterica (serovar Typhi)
Clostridium tetani E88	Schizosaccharomyces pombe
Corynebacterium efficiens YS-314	Shewanella oneidensis MR-1
Deinococcus radiodurans	Shigella flexneri 2a str. 301
Drosophila melanogaster	Sinorhizobium meliloti
Encephalitozoon cuniculi	Staphylococcus aureus MW2
Escherichia coli CFT073	Streptococcus agalactiae 2603V/R
Fugu rubripes	Streptococcus mutans UA159
Fusobacterium nucleatum	Streptococcus pneumoniae R6
Guillardia theta	Streptococcus pyogenes
Haemophilus influenzae Rd	Streptomyces coelicolor A3(2)
Halobacterium sp. NRC-1	Sulfolobus solfataricus
Helicobacter pylori 26695	Synechocystis sp. PCC 6803
Homo sapiens	Thermoanaerobacter tengcongensis
Lactobacillus plantarum WCFS1	Thermoplasma acidophilum
Lactococcus lactis subsp. lactis	Thermosynechococcus elongatus BP-1
Leptospira interrogans 56601	Thermotoga maritima
Listeria innocua	Treponema pallidum
Mesorhizobium loti	Tropheryma whipplei TW08/27
Methanococcus jannaschii	Ureaplasma urealyticum
Methanopyrus kandleri AV19	Vibrio cholerae
Methanosarcina acetivorans str. C2A	Vibrio vulnificus CMCP6
Methanothermobacter thermautotrophicus	Wigglesworthia brevipalpis
Mus musculus	Xanthomonas axonopodis 306
Mycobacterium leprae	Xylella fastidiosa 9a5c
Mycobacterium tuberculosis CDC1551	Yersinia pestis
Mycoplasma genitalium	

Table D.1: All protein sets were obtained from the NCBI.

Appendix E

Publications and Acknowledgements

E.1 Publications

The following chronological list contains peer-reviewed publications that I have authored during the course of my doctoral studies. In the papers where I am listed as first author, the contributions of the other authors are described. All other work was carried out by myself. The papers where I am listed as a secondary author include a summary of my contribution to the work.

1. **Ward, J. J.**, McGuffin, L. J., Buxton, B. F. and Jones, D. T. (2003) Secondary structure prediction with support vector machines, *Bioinformatics*, **19**(13), 1650-1655.

   This paper presents a selection of the results from chapter 2. DTJ provided me with the training set and advice on training the classifier. LJM helped with the benchmarking. BFB provided advice on using SVMs for classification

problems.

2. Jones, D. T. and **Ward, J. J.** (2003) Prediction of disordered regions in proteins from position specific scoring matrices, Proteins, **53**(S6) 573-578.

   The training set and classifer for this paper were developed by DTJ. I was responsible for writing the introduction and generating some of the results from benchmarking.

3. **Ward, J.J.**, Sodhi, J. S., McGuffin, L. J., Buxton, B. F. and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J. Mol. Biol.* **337**(3), 635-45.

   This paper includes results from chapters 3 and 4. DTJ provided the training set. LJM distributed the PSI-BLAST jobs for estimating the frequency of disorder in complete genomes. JSS helped generate the VMD graphics for the 3D protein structures. BFB and DTJ edited the manuscript.

4. **Ward, J.J.**, McGuffin, L.J., Bryson, K., Buxton, B.F. and Jones, D.T. (2004) The DISOPRED prediction of protein disorder server, *Bioinformatics*, **20**(13), 2138-2139.

   The DISOPRED server is based on the java servlet code written by LJM for several other servers from the UCL Bioinformatics Unit. KB provided technical advice. BFB and DTJ edited the manuscript.

5. Sodhi, J.S., Bryson, K., McGuffin, L.J., **Ward, J. J.**, Wernisch, L. and Jones, D. T. (2004) Predicting metal binding sites in low resolution structural models, *J. Mol. Biol.* **342**(1), 307-320.

   I gave advice on training the neural network, provided some matlab code for performing the benchmarking, and edited the manuscript.

6. Sodhi, J. S., McGuffin, L. J., Bryson, K., **Ward, J. J.**, Wernisch, L. and Jones,

D. T. (2004) Automatic prediction of functional site regions in low-resolution protein structures, *IEEE Computational Systems Bioinformatics*, 702-703.

This conference abstract presented a small extension of work carried out in the previous paper by Sodhi et al.

7. **Ward, J. J.**, Sodhi, J. S., Buxton, B. F. and Jones, D. T. (2004) Predicting Gene Ontology annotations from sequence data using kernel-based machine learning algorithms, *IEEE Computational Systems Bioinformatics*, 529-530.

   This extended conference abstract presented some of the preliminary work from chapter 5. The other authors contributed comments on the manuscript.

E.2 Acknowledgements

I would like to begin by thanking Kristina for making me so happy in the years we've spent together, and for being the person who makes all the effort worth while. I would also like to thank my family for all the generous support they have given me, and June and Hamish for their kindess and positivity.

I would also like to thank all the past and present members of the Bioinformatics Unit for their friendship and support over the previous three years. I would specifically like to thank my two supervisors; David Jones for providing a stimulating working environment and for demonstrating the benefits of being pragmatic; Bernard Buxton for being an excellent role model and for providing conscientious comments on this thesis. I would also like to thank Kevin Bryson for providing the entertainment at coffee, and for his constant help and advice. My thanks also go to Liam McGuffin for his benchmarking expertise (nobody does it better...), to Stefano Street for the football chat over a couple of beers, to Tim Ebbels for the enthusiastic scientific discussions, and to Alice Walker-Taylor and Marialuisa Pellegrini-Calace

for the gossip. Special thanks go to Jaz Sodhi for the end-of-day conversations on everything from the meaning of life to (the favourites), "How many pages have you done?- What font?", and for putting up with my heavy right foot on the Pacific drive. Finally, I would like to thank all the other members of the unit; Alistair Coleman, David Corney, Naama Hurwitz, Stefano Lise, Russell Marsden, Chris Pettitt, Ching-Wai Tan and Matthew Trotter. This work was supported by the Medical Research Council (MRC).

Bibliography

Abascal, F., Valencia, A., 2003. Automatic annotation of protein function based on family identification. Proteins 53 (3), 683–692.

Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L., Russell, R. B., 2004. Structure-based assembly of protein complexes in yeast. Science 303, 2026–2029.

Alter, O., Brown, P. O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modelling. Proc. Natl. Acad. Sci. USA 97 (18), 10101–10106.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–10.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl. Acids Res. 25 (7), 3389–3402.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., Yeh, L.-S. L., 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32 Database issue, 115–119.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement in TrEMBLE in 2000. Nucl. Acids Res. 28, 45–48.

Baldi, P., Brunak, S., Frasconi, P., Soda, G., Pollastri, G., 1999. Exploiting the past and future in protein secondary structure prediction. Bioinformatics 15, 937–946.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., Eddy, S. R., 2004. The Pfam protein families database. Nucleic Acids Res 32 Database issue, 138–141.

Beer, M. A., Tavazoie, S., 2004. Predicting gene expression from sequence. Cell 117 (2), 185–198.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., Haussler, D., 2004. Ultraconserved elements in the human genome. Science 304 (5675), 1321–1325.

Bendtsen, J. D., Nielsen, H., von Heijne, G., Brunak, S., 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340 (4), 783–795.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L., 2004. GenBank: update. Nucleic Acids Res 32 Database issue, 23–26.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., 2000. The protein data bank. Nucl. Acids Res. 28, 235–242.

Bishop, C. M., 1995. Neural Networks for Pattern Recognition. Oxford University Press.

Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J.,

Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M., Kim, S. K., 2002. A global analysis of Caenorhabditis elegans operons. Nature 417 (6891), 851–854.

Borel, F., Lohez, O. D., Lacroix, F. B., Margolis, R. L., 2002. Multiple centrosomes arise from tetraploidy checkpoint failure and mitotic centrosome clusters in *p53* and RB pocket protein-compromised cells. Proc. Natl. Acad. Sci. 99, 9819–9824.

Bossemeyer, D., 1995. Protein kinases–structure and function. FEBS Lett 369 (1), 57–61.

Bracken, C., 2001. NMR spin relaxation methods for characterization of disorder and folding in proteins. J Mol Graph Model 19 (1), 3–12.

Branden, C., Tooze, J., 1999. Introduction to Protein Structure, second edition Edition. Garland.

Bray, J. E., Marsden, R. L., Rison, S. C. G., Savchenko, A., Edwards, A. M., Thornton, J. M., Orengo, C. A., 2004. A practical and robust sequence search strategy for structural genomics target selection. Bioinformatics 20 (14), 2288–2295.

Brazma, A., Vilo, J., 2000. Gene expression data analysis. FEBS letters 480, 17–24.

Brenner, S. E., 1999. Errors in genome annotation. Trends Genet 15 (4), 132–133.

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., Dunker, A. K., 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol 55 (1), 104–110.

Brown, M., Grundy, W., Lin, D., Christianini, N., Sugnet, C., Furey, T., Ares, J., Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. USA 97, 262–267.

Burges, C. J. C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Burges, C. J. C., Schölkopf, B., 1997. Improving the accuracy and speed of support vector machines. In: Mozer, M. C., Jordan, M. I., Petsche, T. (Eds.), Advances in Neural Information Processing Systems. Vol. 9. The MIT Press, p. 375. URL `citeseer.ist.psu.edu/burges97improving.html`

Burley, S. K., 2000. An overview of structural genomics. Nature Stuctural Biology 7 (11), 932–934.

Cai, Y.-D., Doig, A. J., 2004. Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition. Bioinformatics 20 (8), 1292–1300, evaluation Studies.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., Apweiler, R., 2003a. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res 13 (4), 662–672.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A., Apweiler, R., 2003b. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res 13 (4), 662–672.

Chandonia, J.-M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S. E., 2004. The ASTRAL Compendium in 2004. Nucleic Acids Res 32 Database issue, 189–192.

Chou, P., Fasman, G., 1974. Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins. Biochemistry 13, 211–222.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., *et al*, 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544.

Crammer, K., Singer, Y., 2001. On the algorithmic implementation of multiclass kernel-based vector machines. Tech. rep., School of Computer Science and Engineering, Hebrew University.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and other Kernel-Based Learning Methods. Cambridge University Press.

Cuff, J. A., Barton, G. J., 1999. Evalution and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 35, 508–519.

Daughdrill, G. W., Hanely, L. J., Dahlquist, F. W., 1998. The C-terminal half of the anti-sigma factor FlgM contains a dynamic equilibrium solution structure favoring helical conformations. Biochemistry 37 (4), 1076–82.

Demarest, S. J., Martinez-Yamout, M., Chung, J., Chen, H., Xu, W., J., D. H., Evans, R. M., Wright, P. E., 2002. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. Nature 415, 549–553.

Donne, D. G., Viles, J. H., Groth, D., Mehlhorn, I., James, T. L., Cohen, F. E., Prusiner, S. B., Wright, P. E., Dyson, H. J., 1997. Structure of the recombinant full-length hamster prion protein PrP(29-231): The N-terminus is highly flexible. Proc. Natl. Acad. Sci. USA 94, 13452–13457.

Downs, T., Gates, K., Masters, A., 2001. Exact simplification of support vector solutions. J. Mach. Learn. Research. 2, 293–297.

DuBay, K. F., Pawar, A. P., Chiti, F., Zurdo, J., Dobson, C. M., Vendruscolo, M., 2004. Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. J Mol Biol 341 (5), 1317–1326.

Duda, R., Hart, P., Stork, D., 2000. Pattern Classification, 2nd Edition. John Wiley and Sons.

Dunker, A., Obradovic, Z., 2001. The protein trinity— linking function and disorder. Nat. Biotechnol. 19 (9), 805–806.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., Obradovic, Z., 2002. Intrinsic disorder and protein function. Biochemistry 41 (21), 6573–6582.

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E., Brown, C., 2000. Intrinsic protein disorder in complete genomes. Genome Informatics 11, 161–171.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis. Cambridge University Press.

Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., Cherry, J., 2002. *Saccharomyces* Genome Database (SGD) provides secondary annotation using the Gene Ontology (GO). Nucl. Acids. Res. 30 (1), 69–72.

Dyson, H. J., Wright, P. E., 2002. Coupling of folding and binding for unstructured proteins. Curr. Opin. Struct. Biol. 12, 54–60.

Efron, B., Tibshirani, R. J., 1993. An introduction to the bootstrap. Chapman and Hall, New York.

Enault, F., Suhre, K., Abergel, C., Poirot, O., Claverie, J.-M., 2003. Annotation of bacterial genomes using improved phylogenomic profiles. Bioinformatics 19, Suppl. 1, 105–107.

Fondon, J. W. r., Garner, H. R., 2004. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A 101 (52), 18058–18063.

Fraser, A. G., Marcotte, E. M., 2004. A probabilistic view of gene function. Nat Genet 36 (6), 559–564.

Frishman, D., Argos, P., 1995. Knowledge-based protein secondary structure assignment. Proteins 23 (4), 566–579.

Furey, T. S., Christianini, N., Duffy, N., Bednarski, D. W., Shummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissues using microarray expression data. Bioinformatics 16, 906–914.

Garnier, J., Osguthorpe, D., Robson, B., 1978. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol 120, 97–120.

Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A. *et al.*, 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415 (6868), 141–147.

Gene Ontology Consortium, 2000. Gene Ontology: Tool for the unification of biology. Nature Genetics 25, 25–29.

Gene Ontology Consortium, 2001. Creating the Gene Ontology resource: Design and implementation. Genome Research 11, 1425–1433.

Gerstein, M., Echols, N., 2004. Exploring the range of protein flexibility, from a structural proteomics perspective. Curr Opin Chem Biol 8 (1), 14–19.

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T., Charbit, A., Chetouani, F., Couve, E. *et* al.., 2001. Comparative genomics of Listeria species. Science 294 (5543), 849–852.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S. G., 1996. Life with 6000 genes. Science 274, 546–567.

Goh, C.-S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G. T., Zhao, H., Gerstein, M., 2004. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. J Mol Biol 336 (1), 115–130.

Golub, G., van Loan, C., 1996. Matrix computations, 3rd Edition. The Johns Hopkins University Press, London.

Granzier, H., Helmes, M., Trombitas, K., 1996. Nonuniform elasticity of titin in cardiac myocytes: a study using immunoelectron microscopy and cellular mechanics. Biophys J 70 (1), 430–442.

Graur, D., Li, W.-H., 2000. Fundamentals of Molecular Evolution, 2nd Edition. Sinauer Associates, Sunderland, MA.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., Baker, D., 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. 331, 281–99.

Gu, X., Huang, W., Xu, D., Zhang, H., Apr 2005. GeneContent: software for whole-genome phylogenetic analysis. Bioinformatics 21 (8), 1713–1714.

Gunasekaran, K., Tsai, C., Kumar, S., Zanuy, D., Nussinov, R., 2003. Extended disordered proteins: targeting function with less scaffold. Trends Biochem. Sci 28, 81–85.

Gundersen, G. G., Cook, T. A., 1999. Microtubules and signal transduction. Curr. Opin. Cell Biol. 11, 81–94.

Hanley, J. A., McNeil, B. J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Hanley, J. A., McNeil, B. J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839–843.

Hastie, T., Tibshirani, R., 1998. Classification by pairwise coupling.

Haupt, S., Berger, M., Goldberg, Z., Haupt, Y., 2003. Apoptosis - the p53 network. J Cell Sci 116 (Pt 20), 4077–4085.

Hegyi, H., Gerstein, M., 1999. The relationship between structure and function: a comprehensive survey with application to the yeast genome. J. Mol. Biol. 288, 147–164.

Holm, L., Sander, C., 1998. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 14 (5), 423–429.

Hsu, C.-W., Lin, C.-J., 2002. A comparison on methods for multi-class support vector machines. IEEE Transactions on Neural Networks 13, 415–425.

Hua, S., Sun, Z., 2001. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. J. Mol. Biol. 308, 397–407.

Hulo, N., Sigrist, C. J. A., Le Saux, V., Langendijk-Genevaux, P. S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A., 2004. Recent improvements to the PROSITE database. Nucleic Acids Res 32 Database issue, 134–137.

Humphrey, W., Dalke, A., Schulten, K., 1996. Vmd— Visual Molecular Dynamic. J. Mol. Graph 14, 33–38.

Huson, D. H., Steel, M., 2004. Phylogenetic trees based on gene content. Bioinformatics 20 (13), 2044–2049, evaluation Studies.

Huynen, M., Snel, B., Lathe, W. r., Bork, P., 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 10 (8), 1204–1210.

Huynen, M. A., Snel, B., von Mering, C., Bork, P., 2003. Function prediction and protein networks. Curr Opin Cell Biol 15 (2), 191–198.

Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., Dunker, K. A., 2002. Intrinsic disorder in cell-signalling and cancer-associated proteins. J. Mol. Biol. 323, 573–584.

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., Dunker, A. K., 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32 (3), 1037–1049, evaluation Studies.

Jaakkola, T., Diekhans, M., Haussler, D., 2000. A discriminative framework for detecting remote protein homologies. J. Comp. Biol. 7 (1-2), 95–114.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., Gerstein, M., 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302 (5644), 449–453, evaluation Studies.

Jensen, L. J., Gupta, R., Staerfeldt, H.-H., Brunak, S., 2003. Prediction of human protein function according to gene ontology categories. Bioinformatics 19 (5), 635–642.

Joachims, T., 1999. Making large-scale SVM learning practical.

Jones, D., Swindells, M., 2002. Getting the most from PSI-BLAST. Trends in Biochem. Sci. 27, 161–164.

Jones, D. T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 196–202.

Jones, D. T., 2001. Predicting novel protein folds by using FRAGFOLD. Proteins Suppl 5, 127–132.

Jones, D. T., Ward, J. J., 2003. Prediction of disordered regions in proteins from position specific scoring matrices. Proteins 53 (S6), 573–578.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Kall, L., Krogh, A., Sonnhammer, E. L. L., 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338 (5), 1027–1036.

Kalodimos, C. J., Biris, N., Bonvin, A., Levandoski, M. M., Guennuegues, M., Boelens, R., Kaptein, R., 2004. Structure and flexibility adaptation in nonspecific and specific protein-dna complexes. Science 305, 386–389.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., Jan 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res 32 (Database issue), 277–280.

Karlin, S., Altschul, S., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. 87 (6), 2264–8.

Karplus, K., Barrett, C., Hughey, R., 1998. Hidden markov models for detecting remote protein homologies. Bioinformatics 14, 846–856.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E. S., 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423 (6937), 241–254.

Kersey, P. J., Morris, L., Hermjakob, H., Apweiler, R., 2003. Integr8: enhanced interoperability of European molecular biology databases. Methods Inf Med 42 (2), 154–160.

Khanna, K. K., Jackson, S. P., 2001. DNA double-strand breaks: signaling, repair and the cancer connection. Nature Genetics 27, 247–254.

Kondor, R. I., Lafferty, J. D., 2002. Diffusion kernels on graphs and other discrete input spaces. ICML, 315–322.

Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L., Baker, D., 2004. Computational redesign of protein-protein interaction specificity. Nat Struct Mol Biol 11 (4), 371–379.

Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I., Wright, P. E., 1996. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. Proc Natl Acad Sci U S A 93 (21), 11504–11509.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., Haussler, D., 1994. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235 (5), 1501–1531.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., Baker, D., 2003. Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364–1368.

Kyogoku, Y., Fujiyoshi, Y., Shimada, I., Nakamura, H., Tsukihara, T., Akutsu, H.,

Odahara, T., Okada, T., Nomura, N., 2003. Structural genomics of membrane proteins. Acc Chem Res 36 (3), 199–206.

Kyrpides, N., 1999. Genomes OnLine Database (GOLD): a monitor of complete and ongoing genome projects world wide. Bioinformatics 15, 773–774.

Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., Noble, W. S., 2004. A statistical framework for genomic data fusion. Bioinformatics 20 (16), 2626–2635.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.*, 2001. Initial sequencing and analysis of the human genome. Nature 409 (6822), 860–921.

Laskowski, R. A., Watson, J. D., Thornton, J. M., 2003. From protein structure to biochemical function? J Struct Funct Genomics 4 (2-3), 167–177.

Lee, J.-H., Lin, C.-J., 2000. Automatic model selection for support vector machines. Tech. rep., Department of Computer Science and Information Engineering, National Taiwan University.
URL `http://www.csie.ntu.edu.tw/ cjlin/papers/modelselect.ps.gz`

Lee, W., Harvey, T. S., Yin, Y., Yau, P., Litchfield, D., Arrowsmith, C. H., 1995. Solution structure of the tetrameric minimum transforming domain of p53. Nat. Struct. Biol. 81, 877–890.

Lesk, A. M., 1997. CASP2 : Report on *ab initio* predictions. Proteins Suppl 1, 151–166.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., Bork, P., 2004. SMART 4.0: towards genomic data integration. Nucleic Acids Res 32 (Database issue), 142–144.

Li, X., Romero, P., Rani, M., Dunker, A. K., Obradovic, Z., 1999. Predicting protein disorder for N-, C-, and internal regions. Genome Informatics 10, 30–40.

Linding, R., Jensen, L., Diella, F., Bork, P., Gibson, T., Russell, R., 2003. Protein disorder prediction. Implications for structural proteomics. Structure 11 (11), 1453–1459.

Lise, S., Jones, D., 2004. Sequence patterns associated with disordered regions in proteins. Proteins 58 (1), 144–150.

Liu, J., Tan, H., Rost, B., 2003. Loopy proteins appear conserved in evolution. J. Mol. Biol. 322, 53–64.

Lodhi, H., Shawe-Taylor, J., Cristianini, N., Watkins, C. J. C. H., 2000. Text classification using string kernels. In: NIPS. pp. 563–569.
URL `citeseer.ist.psu.edu/lodhi02text.html`

Mackay, D., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Marcotte, C. J. V., Marcotte, E. M., 2002. Predicting functional linkages from gene fusions with confidence. Applied Bioinformatics 1 (2), 93–100.

Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., Eisenberg, D., 1999a. Detecting protein function and protein-protein interactions from genome sequences. Science 285, 751–753.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., Eisenberg, D., 1999b. A combined algorithm for genome-wide prediction of protein function. Nature 402, 83–86.

Marcotte, E. M., Xenarios, I., van der Bliek, A. M., Eisenberg, D., 2000. Localizing

proteins in the cell form their phylogenetic profiles. Proc. Natl. Acad. Sci. USA 97 (22), 12115–12120.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C., Thornton, J. M., 1998. Protein folds and functions. Structure 6 (7), 875–884.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, *et al.*., 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31 (1), 374–378.

Mazza, C., Segref, A., Mattaj, I. W., Cusack, S., 2002. Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. EMBO J. 21 (20), 5548–57.

McGuffin, L. J., Bryson, K., Jones, D. T., 2000. The psipred protein structure prediction server. Bioinformatics 16, 404–405.

McGuffin, L. J., Bryson, K., Jones, D. T., 2001. What are the baselines for protein fold recognition? Bioinformatics 17, 63–72.

McGuffin, L. J., Jones, D. T., 2002. Benchmarking secondary structure prediction for fold recognition. Proteins 17, 63–72.

McGuffin, L. J., Jones, D. T., 2003. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 19 (7), 874–881, evaluation Studies.

Meiler, J., Baker, D., 2003. Coupled prediction of protein secondary and tertiary structure. Proc. Natl. Acad. Sci. 100 (21), 12105–12110.

Melamud, E., Moult, J., 2003. Evaluation of disorder predictions in CASP5. Proteins 53 (S6), 561–565.

Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., Ruepp, A., 2004. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res 32 Database issue, 41–44.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller., K.-R., 1999. Fisher discriminant analysis with kernels. In: Hu, Y.-H., Larsen, J., Wilson, E., Douglas, S. (Eds.), Neural Networks for Signal Processing IX. IEEE. pp. 41–48.

Morik, K., Brockhausen, P., Joachims, T., 1999. Combining statistical learning with a knowledge-based approach— A case study in intensive care monitoring. In: International Conference on Machine Learning (ICML).

Mott, R., Schultz, J., Bork, P., Ponting, C. P., 2002. Predicting protein cellular localization using a domain projection method. Genome Res 12 (8), 1168–1174.

Murzin, A., Brenner, S., Hubbard, T., Chothia, C., 1995. SCOP— a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Nakashima, H., Nishikawa, K., 1992. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Lett 303 (2-3), 141–146.

Nakayama, K., Hatakeyama, S., Nakayama, K., 2001. Regulation of the cell cycle at the G1-S transition by proteolysis of cyclin E and p27Kip1. Biochem. Biophys. Res. Commun 282 (4), 853–860.

Needleman, S. B., Wunsch, C. D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Nephew, K. P., Huang, T. H., 2003. Epigenetic gene silencing in cancer initiation and progression. Cancer Lett. 190 (2), 125–133.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., Dunker, A. K., 2003. Predicting intrinsic disorder from amino acid sequence. Proteins 53 (S6), 566–572.

Orlov, Y. L., Potapov, V. N., 2004. Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res 32 (Web Server issue), 628–633.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., Chothia, C., 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284 (4), 1201–1210.

Park, K.-J., Kanehisa, M., 2003. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19 (13), 1656–1663, evaluation Studies.

Parker, D., Rivera, M., Zor, T., Henrion-Caude, A., Radhakrishnan, I., Kumar, A., Shapiro, L. H., Wright, P. E., Montminy, M., Brindle, P. K., 1999. Role of secondary structure in discrimination between constitutive and inducible activators. Mol. Cell. Biol. 19, 5601–5607.

Pavlidis, P., Weston, J., Cai, J., Stafford Noble, W., 2002. Learning gene functional classifications from multiple data types. J. Comp. Biol. 9 (2), 401–411.

Pazos, F., Sternberg, M. J. E., 2004. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci U S A 101 (41), 14754–14759.

Pearson, W. R., 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183, 63–98.

Pearson, W. R., Lipman, D. J., 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85 (8), 2444–2448.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., Yeates, T. O., 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA 96, 4285–4288.

Perneger, T. V., 1998. What's wrong with Bonferroni adjustments. BMJ 316 (7139), 1236–1238.

Phizicky, E., Bastiaens, P., Zhu, H., Snyder, M., Fields, S., 2003. Protein analysis on a proteomic scale. Nature 422, 208–215.

Platt, J., 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola, A., Bartlett, P., Schoelkopf, B., Schuurmans, D. (Eds.), Advances in Large Margin Classifiers. pp. 61–74.
URL `citeseer.ist.psu.edu/platt99probabilistic.html`

Platt, J. C., 1999. Fast training of support vector machines using sequential minimal optimisation. In: Schölkopf, B., Burges, C. J. C., Smola, A. J. (Eds.), Advances in Kernel Methods— Support Vector Learning. MIT press, pp. 185–208.

Platt, J. C., Christianini, N., Shawe-Taylor, J., 2000. Large margin DAGs for multiclass classification. In: Solla, S., Leen, T. K., Müller, K. R. (Eds.), Advances in Neural Information Processing systems. Vol. 12. MIT press, pp. 547–553.

Pollastri, G., Przybylski, D., Rost, B., Baldi, P., 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47, 228–235.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., 1988. Numerical recipes in C, the art of scientific computing. Cambridge University Press.

Prilusky, J., Zeev-Ben-Mordehai, T., Rydberg, E., Felder, C., Silman, I., Sussman, J. L., 2003. Foldindex.
URL `http://bioportal.weizmann.ac.il/fldbin/findex`

Qian, N., Sejnowski, T. J., 1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202, 865–884.

Radhakrishnan, I., Perez-Alvarado, G. C., Parker, D., Dyson, H. J., Montminy, M. R., Wright, P. E., 1997. Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. Cell 91 (6), 741–752.

Rees, D. G., 1995. Essential Statistics, 3rd Edition. Chapman and Hall.

Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: Proc. of the IEEE Intl. Conf. on Neural Networks. pp. 586–591.

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., Baker, D., 2004. Protein structure prediction using Rosetta. Methods Enzymol 383, 66–93.

Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J., Dunker, A. K., 1997. Identifying disordered regions in proteins from amino acid sequences. IEEE Int. Conf. Neural Netw. 1, 90–95.

Romero, P., Obradovic, Z., Li, X., Garner, E., Brown, C., Dunker, A., 2001. Sequence complexity and disordered proteins. Proteins 42, 38–48.

Rost, B., 1996. PHD: predicting 1D protein structure by profile-based neural networks. Meth. Enzym. 266, 525–539.

Rost, B., 2001a. Protein structure prediction continues to rise. J. Struct. Biol. 134, 204–218.

Rost, B., 2001b. Twilight zone of protein sequence alignments. Protein Eng. 134, 204–218.

Rost, B., 2002. Enzyme function less conserved than anticipated. J Mol Biol 318 (2), 595–608.

Rost, B., Eyrich, V. A., 2001. EVA: Large-scale analysis of secondary structure prediction. Proteins 5, 192–199.

Rost, B., Sander, C., 1993. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol 232, 584–99.

Rost, B., Sander, C., 2000. Third generation prediction of secondary structures. In: Webster, D. (Ed.), Methods in Molecular Biology. Humana Press in., Totowa, NJ, Ch. 5, pp. 71–96.

Sayle, R. A., Milner-White, E. J., 1995. RASMOL: biomolecular graphics for all. Trends Biochem Sci 20 (9), 374.

Schwartz, R. M., Dayhoff, M. O., 1978. Origins of prokaryotes, eukaryotes, mitochondria and chloroplasts. Science 199, 395–403.

Sherman, F., 1997. Yeast genetics. In: Meyers, R. A. (Ed.), The Encyclopedia of Molecular Biology and Molecular Medicine. Vol. 6. VCH Pub. Weinheim, Germany, pp. 302–325.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., Golub, T. R., 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8 (1), 68–74.

Smith, T. F., Waterman, M., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 227–258.

Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L., Jones, D. T., 2004a. Predicting metal-binding site residues in low-resolution structural models. J Mol Biol 342 (1), 307–320.

Sodhi, J. S., McGuffin, L. J., Bryson, K., Ward, J. J., Wernisch, L., Jones, D. T., 2004b. Automatic prediction of functional site regions in low-resolution protein structures. In: IEEE Computational Systems Bioinformatics. pp. 702–703.

Spolar, R. S., Record, M. T., 1994. Coupling of local folding to site-specific binding of proteins to DNA. Science 263, 777–784.

Stryer, L., 1995. Biochemistry, 4th Edition. W. H. Freeman & Co. New York.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., Natale, D. A., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4 (1), 41.

Tipping, M. E., 2001. Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res., 211–244.

Todd, A., Orengo, C., Thornton, J., 2001. Evolution of function in protein superfamilies from a structural perspective. J. Mol. Biol. 307, 1113–1143.

Tompa, P., 2002. Intrinsically unstructured proteins. Trends Biochem Sci 27 (10), 527–533.

Tompa, P., 2003. Intrinsically unstructured proteins evolve by repeat expansion. Bioessays 25 (9), 847–855.

Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J., 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403, 623–627.

Uversky, V., Gillespie, J., Fink, A., 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41 (3), 415–427.

Vapnik, V., 1998. Statistical Learning Theory. John Wiley and Sons, New York.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.*, 2001. The sequence of the human genome. Science 291 (5507), 1304–1351.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., Smith, H. O., 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304 (5667), 66–74.

Vert, J.-P., 2002. A tree kernel to analyse phylogenetic profiles. Bioinformatics 18 (Suppl. 1), S276–S284.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., Bork, P., 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33 Database Issue, 433–437.

Vucetic, S., Brown, C. J., Dunker, A. K., Obradovic, Z., 2003. Flavors of protein disorder. Proteins 52 (4), 573–584.

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., Cortese, M. S., Lawson, J. D., Brown, C. J., Sikes, J. G., Newton, C. D., Dunker, A. K., Jan 2005. DisProt: a database of protein disorder. Bioinformatics 21 (1), 137–140.

Ward, J. J., Bryson, K., McGuffin, L. J., , Buxton, B. F., Jones, D. T., 2004a. The DISOPRED prediction of protein disorder server. Bioinformatics 20 (13), 2138–2139.

Ward, J. J., McGuffin, L. J., Buxton, B. F., Jones, D. T., 2003. Secondary structure prediction with support vector machines. Bioinformatics 19 (13), 1650–1655.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., Jones, D. T., 2004b. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. 337, 635–645.

Webb, E. C., 1992. Enzyme Nomenclature. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, Academic Press, New York.

Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C., Struhl, K., 1990. Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. Nature 347, 575–578.

Weisstein., E. W., 2004. Binomial distribution.
URL `http://mathworld.wolfram.com/BinomialDistribution.html`

Weston, J., Watkins, C., 1998. Multi-class support vector machines. Tech. rep., Royal Holloway.

Wilson, C. A., Kreychman, J., Gerstein, M., 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297 (1), 233–249.

Winzeler, E. A. *et al.*., 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 285, 901–906.

Wodak, S. J., Mendez, R., 2004. Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 14 (2), 242–249.

Wolf, Y. I., Grishin, N. V., Koonin, E. V., 2000. Estimating the number of protein folds and families from complete genome data. J Mol Biol 299 (4), 897–905.

Wright, P. E., Dyson, H. J., 1999. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331.

Xenarios, I., Rice, D., Salwinski, L., Baron, M., Marcotte, E., Eisenberg, D., 2000. DIP: The database of interacting proteins. Nucl. Acids Res. 28, 289–91.

Xia, Y., Levitt, M., 2004. Simulating protein evolution in sequence and structure space. Curr Opin Struct Biol 14 (2), 202–207.

Zemla, A., Venclovas, C., Fidelis, K., Rost, B., 1999. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 34, 220–223.

Zhang, Z., Harrison, P. M., Liu, Y., Gerstein, M., 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13 (12), 2541–2558.

Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., Müller, K.-R., 2000. Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics 16, 799–807.